

## ABSTRACT

Title of Document: GENOMIC STUDIES OF THE EVOLUTION OF  
HAPTOPHYTES AND DINOFLAGELLATES WITH  
EMPHASIS ON THE CHROMALVEOLATE  
HYPOTHESIS

Maria Virginia Sanchez Puerta, Doctor of Philosophy, 2006

Directed By: Professor Charles F. Delwiche, Department of Cell Biology  
and Molecular Genetics

All photosynthetic eukaryotes rely, partially or totally, on their plastids to live. The plastids, which ultimately are highly modified cyanobacteria, were acquired through a process of primary, secondary, or tertiary endosymbiosis. Four photosynthetic lineages, including haptophytes, dinoflagellates, cryptophytes, and heterokonts, contain secondary plastids with chlorophyll *c* as a main photosynthetic pigment. These four lineages were grouped together, along with their heterotrophic relatives, on the basis of their pigmentation and called chromalveolates by Cavalier-Smith. However, the phylogenetic relationships among these algae are unknown and the chromalveolate hypothesis remains very controversial. This study focuses on increasing the amount of genomic data from a poorly studied chromalveolate lineage, the haptophytes, and understanding plastid evolution in chromalveolates. Both the chloroplast and mitochondrial genomes of the haptophyte *Emiliania huxleyi* were sequenced and examined to describe basic genomic properties, as well as perform comparative studies. Phylogenetic analyses, including data

acquired from haptophytes, support a monophyletic chl c containing plastid clade derived from the red algae, after the divergence of Cyanidiales, with the cryptophyte plastid basal or sister to the haptophyte plastid. In addition, phylogenetic analyses using mitochondrial data suggest a relationship of haptophytes and cryptophytes. The chromalveolate clade as a whole is not recovered nor rejected by the data. Analysis of an EST project from the heterotrophic dinoflagellate *Cryptothecodinium cohnii* indicates that *C. cohnii* is not only derived from a photosynthetic ancestor, but very likely retains a non-photosynthetic plastid. Analyses of putative gene function suggest that heme biosynthesis, non-mevalonate isoprenoid biosynthesis, amino-acid metabolism, and Fe-S cluster assembly may occur in the plastid. These observations are also consistent with the chromalveolate hypothesis, which proposes that several major groups of eukaryotes, including alveolates, haptophytes, cryptophytes, and heterokonts, may form a monophyletic group with a photosynthetic common ancestor, and that nonphotosynthetic members are secondarily so.



GENOMIC STUDIES OF THE EVOLUTION OF HAPTOPHYTES AND  
DINOFLAGELLATES WITH EMPHASIS ON THE CHROMALVEOLATE  
HYPOTHESIS

By

Maria Virginia Sanchez Puerta

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park, in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2006

Advisory Committee:  
Professor Charles F. Delwiche, Chair  
Professor Elisabeth Gantt  
Professor Diane K. Stoecker  
Professor Steve Wolniak  
Professor Charles Mitter

© Copyright by  
Maria Virginia Sanchez Puerta  
2006

## Dedication

To my grandmother, for her infinite love and care, and to my parents, who mean everything to me.

## Acknowledgements

I would like to thank my advisor, Charles Delwiche, whose invaluable advice and encouragement helped me develop my research project, improve my overall education, and identify career goals. I am also grateful to Tsvetan Bachvaroff for guiding me during my initial steps in the lab to conduct this project and for passing on his passion for protist evolution, as well as sharing productive and exciting discussions. I would like to specially thank my dear labmate, John Hall, who accompanied and listened to me in my happy days, and put up with me during the crazy, stressed ones, making the journey much more enjoyable. I also express my gratitude to the members of my committee for their useful suggestions and discussions on this research: Charles Mitter, Elisabeth Gantt, Diane Stoecker, and Steve Wolniak. I thank Micah S. Durnthorn for reading and commenting on several sections of this dissertation. I want to specially acknowledge my undergraduate mentor from Argentina, Patricia I. Leonardi, who supervised me during my first lab experiences, and engaged me in algal research with her contagious enthusiasm. She inspired and advised me in every step of the way, and I look up to her not only for being an excellent researcher but a great person.

I want to most sincerely thank my family, who are most important to me and taught me the value of education and perseverance. My dad helped me keep an open mind and see things in perspective. My mom stood by me giving me constant support and challenging me to be a better person. My sister encouraged me to pursue a PhD in the USA and continuously rooted for me. I also want to thank Valdemar, for his support, patience, and understanding during this process. Last but not least, I want to

thank my beloved friends Facundo, Susana, Henry, Patricia, Edgar, Maximo, Faten, Jaquan, Laura, and Shaun who brought much happiness to my life and helped me in everyday hassle, making my experience in College Park an unforgettable one.

## Table of Contents

Table of Contents .....	v
List of Tables .....	viii
List of Figures .....	ix
Chapter I – General Introduction .....	1
Scientific Background .....	1
Diversity of life .....	1
Origin of photosynthetic eukaryotes .....	3
Chromalveolates .....	8
Haptophytes .....	11
Dinoflagellates .....	15
Cryptophytes .....	18
Heterokonts or Stramenopiles .....	20
Hypotheses of plastid evolution in chromalveolates .....	21
Objectives and significance .....	24
Chapter II – The Complete Mitochondrial Genome Sequence of the Haptophyte	
<i>Emiliania huxleyi</i> and its Relation to Heterokonts .....	28
Abstract .....	28
Introduction .....	29
Materials and Methods .....	30
Culture of <i>E. huxleyi</i> and mtDNA isolation .....	30
Cloning and DNA sequencing .....	31
Data analysis .....	32
Phylogenetic trees .....	32
Results and Discussion .....	35
Overall organization of <i>E. huxleyi</i> mtDNA .....	35
Gene content .....	37
A novel feature in mtDNA: adenine methyltransferase .....	40
Codon usage .....	43
Transfer RNAs .....	45
Phylogenetic analysis based on mitochondrial genes .....	46
Haptophyte evolution .....	51
Chapter III – The Complete Plastid Genome Sequence of the Haptophyte <i>Emiliania</i>	
<i>huxleyi</i> : a Comparison to Other Plastid Genomes .....	56
Abstract .....	56
Introduction .....	57
Materials and Methods .....	60
Culture of <i>E. huxleyi</i> and cpDNA isolation .....	60
Library construction and DNA sequencing .....	62

Data analysis-----	62
Cladistic analysis-----	63
Results and Discussion -----	63
Overall organization of <i>E. huxleyi</i> cpDNA-----	63
Codon usage and transfer RNA genes -----	66
Gene content in <i>E. huxleyi</i> cpDNA-----	69
Comparison of gene content and function among all plastid genomes-----	71
Cluster analysis-----	74
Parsimony analysis of presence and absence of genes -----	76
Chapter IV - Phylogenetic Signal vs. Noise in Plastid Genomic Data and the Evolution of Chlorophyll c Containing Plastids -----	81
Abstract-----	81
Introduction -----	82
Materials and Methods -----	86
Sequence acquisition and alignment -----	86
Phylogenetic analyses -----	88
Approximately Unbiased (AU) tests -----	89
Results-----	89
Individual gene analyses -----	89
Concatenated gene analyses-----	103
Approximately Unbiased (AU) test -----	109
Discussion-----	110
Individual gene analyses -----	111
Reliability of different datasets and analytical methods -----	112
Relationships among glaucophyte, red, and green plastids and their host cells	116
Other observations: green plastid phylogeny -----	117
Red algal plastids are monophyletic or paraphyletic?-----	118
Are chl c containing plastids monophyletic?-----	119
Are chl c containing host cells monophyletic?-----	119
Relationships among chl c containing plastids-----	122
Conclusions -----	123
Chapter V – The Heterotrophic Dinoflagellate <i>Cryptothecodinium cohnii</i> Descends from a Plastid-bearing Ancestor, Suggesting an Earlier Acquisition of Plastids ----	125
Abstract-----	125
Introduction -----	126
Materials and Methods -----	130
Strain and cultivation -----	130
Library construction, sequence and analysis -----	130
Identification of plastid-associated genes -----	130
Phylogenetic analyses -----	132
Targeting prediction -----	132
Results-----	133
Putative plastid-associated genes in <i>C. cohnii</i> -----	133
Phylogenetic analyses -----	135
Targeting signal prediction-----	140
Discussion-----	142

Evidence for past endosymbiosis-----	142
Rubisco-----	143
Models of evolution-----	144
Origin of plastid-derived genes in <i>C. cohnii</i> -----	147
Predicting protein targeting -----	148
Plastid-related metabolism in <i>C. cohnii</i> and comparison to other algae -----	150
Histones -----	156
Conclusions -----	157
Chapter VI – Conclusions and Future Directions -----	159
Conclusions -----	159
Future Directions -----	166
References -----	169



## List of Tables

I.1. Description of plastids in several lineages of eukaryotes -----	9
II.1. Sources of sequences used in phylogenetic analyses -----	34
II.2. Genes identified in <i>E. huxleyi</i> mtDNA -----	38
II.3. Codon usage in the mitochondrial genome of <i>E. huxleyi</i> -----	44
II.4. Bootstrap support values in Maximum Likelihood analyses -----	47
III.1. List of plastid genomes from different photosynthetic eukaryotes -----	58
III.2. List of genes used for the Maximum Parsimony analysis -----	61
III.3. Codon usage in the chloroplast genome of <i>E. huxleyi</i> -----	67
III.4. List of genes encoded in the plastid genome of <i>E. huxleyi</i> -----	68
IV.1. Taxonomy of photosynthetic eukaryotes and taxon sampling -----	85
IV.2. Published sequences used in the phylogenetic analyses -----	87
IV.3. Individual phylogenetic analyses based on plastid-associated genes, including dinoflagellates -----	90
IV.4. Individual phylogenetic analyses based on plastid-associated genes, excluding dinoflagellates -----	99
IV.5. Datasets based on plastid-associated genes used for phylogenetic analyses -	104
V.1. Putative plastid-associated genes in <i>C. cohnii</i> -----	134

## List of Figures

I.1. Three domains of life -----	2
I.2. Acquisition of plastids in photosynthetic eukaryotes -----	5
I.3. Diagram of relationships among main eukaryotic lineages -----	7
I.4. Diagram of a cell of the coccolithophorid <i>Emiliana huxleyi</i> -----	13
I.5. Proposed models of chromalveolate evolution -----	22
II.1. Physical map of the <i>E. huxleyi</i> mitochondrial genome -----	36
II.2. Phylogenetic tree based on the <i>dam</i> gene -----	41
II.3. Individual gene phylogenetic analyses -----	48
II.4. Maximum likelihood tree based on the genes <i>cob</i> , <i>cox1</i> , <i>cox2</i> , and <i>cox3</i> -----	52
II.5. Phylogenetic analyses based on <i>cob</i> , <i>cox1</i> , <i>cox2</i> , and <i>cox3</i> -----	53
II.6. Phylogenetic analyses based on 13 mitochondrial genes -----	54
III.1. Physical map of the <i>E. huxleyi</i> plastid genome -----	64
III.2. Venn diagram comparing the protein-coding gene content of six plastid genomes -----	70
III.3. Conserved gene clusters in eight plastid genomes and a cyanobacterium -----	75
III.4. Maximum Parsimony tree based on presence or absence of 261 plastid genes, using Camin-Sokal parsimony -----	77
III.5. Maximum Parsimony tree based on presence or absence of 261 plastid genes, using Fitch parsimony -----	78
IV.1. Maximum likelihood analyses based on individual plastid genes, including dinoflagellates -----	92

IV.2. Maximum likelihood analyses based on individual plastid genes, excluding dinoflagellates -----	100
IV.3. Phylogenetic analyses based on concatenated plastid-associated genes, excluding the dinoflagellate -----	105
IV.4. Phylogenetic analyses based on concatenated plastid-associated genes, including the dinoflagellate -----	106
IV.5. Relative codon usage based on the 24-plastid gene dataset -----	113
IV.6. Diagram of evolutionary hypotheses of photosynthetic eukaryotes and their plastids -----	120
V.1. Flow diagram of the procedure to identify putative plastid-derived genes ----	131
V.2. Phylogenetic analyses of putative plastid-targeted genes in <i>C. cohnii</i> -----	136
V.3. Phylogenetic analyses of putative plastid-associated genes in <i>C. cohnii</i> -----	137
V.4. Phylogenetic analyses of putative plastid-associated genes in <i>C. cohnii</i> -----	138
V.5. Phylogenetic analyses of putative plastid-associated genes in <i>C. cohnii</i> -----	139
V.6. Kyte-Doolittle hydropathy plots -----	141
V.7. Models of plastid acquisition in Alveolates -----	146
V.8. Plastid-related metabolism in <i>C. cohnii</i> -----	152
VI.1. Proposed model of chromalveolate evolution -----	161

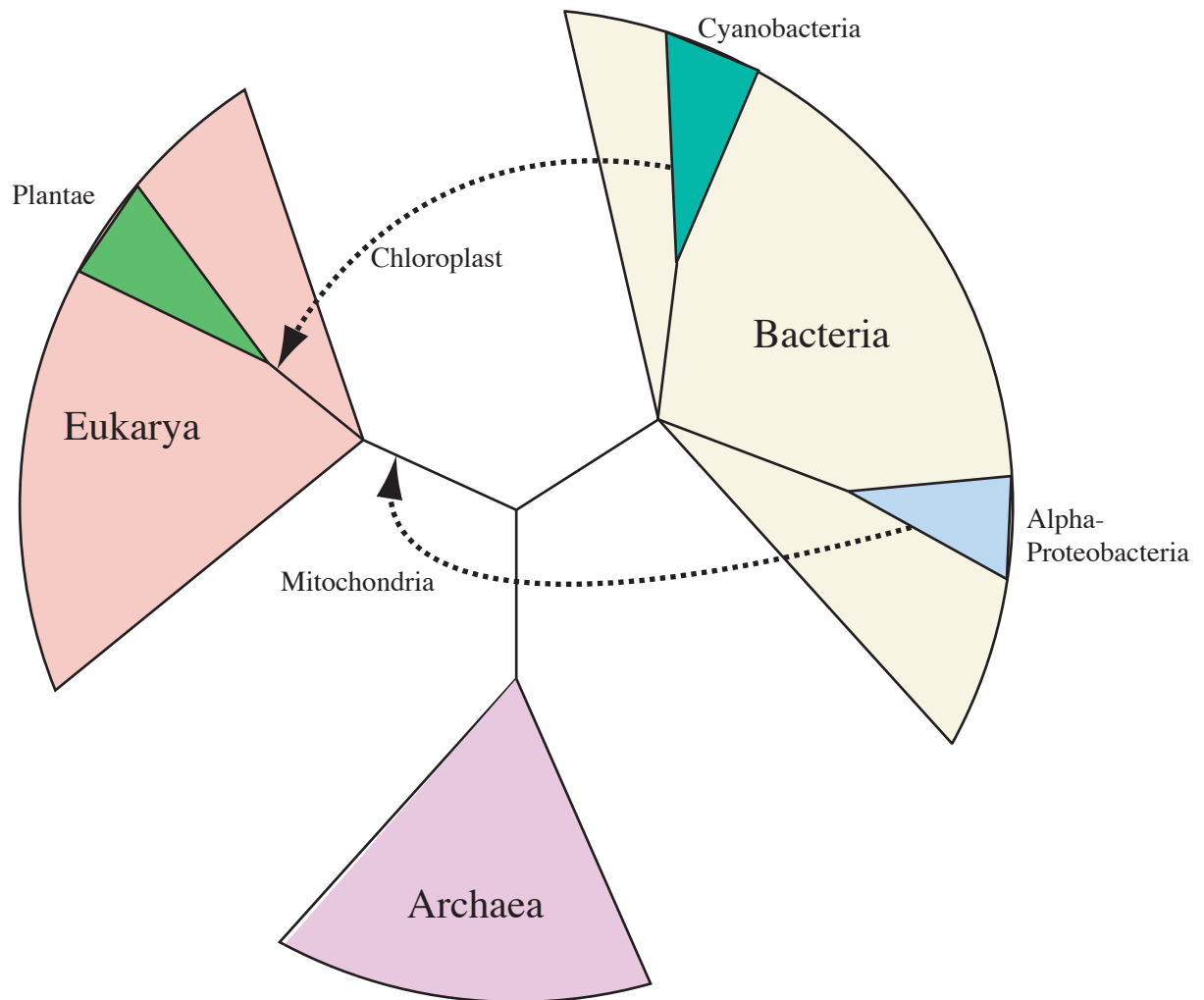
# Chapter I – General Introduction

## *Scientific Background*

### **Diversity of life**

Life on Earth has existed for about 3-4 billion years, evolving into a variety of forms and shapes (Brasier et al. 2002; Schopf et al. 2002; Falkowski et al. 2004). Living organisms have been classified at the molecular level in three domains, namely Archaea, Bacteria, and Eukarya (Figure I.1) (Woese and Fox 1977; Woese et al. 1990). The first two domains have a prokaryotic cell structure, and fossils from them are the oldest known, suggesting that LUCA (last universal common ancestor) was probably prokaryotic (Baldauf et al. 2004). The oldest eukaryotic fossils are 1.8 billion years old (Knoll 1992; Xiao et al. 1998; Baldauf et al. 2004). Distinction of these three major lineages was possible by the use of phylogenetic inference (Woese and Fox 1977), but interdomain relationships are not fully resolved, partly because phylogenetic reconstruction has been confounded by the effect of lateral gene transfer across domains. The root of the tree of life most probably lies within the Bacteria, leaving Archaea and Eukaryotes as sister taxa (Gogarten et al. 1989; Iwabe et al. 1989; Woese et al. 1990), but other interpretations have been postulated, including the “ring of life” hypothesis, in which eukaryotes arose from the fusion of bacterial and archaeal prokaryotes (Rivera and Lake 2004). Archaea is a poorly studied group characterized by single cells, many of which inhabit extreme environments but with representatives in more mesic environments as well. Bacteria are the most abundant

Figure I.1. Three domains of life. Eukaryotes acquired mitochondria and chloroplasts by engulfing an alpha-proteobacterium and a cyanobacterium, respectively.



organisms on Earth with highly variable features and they can be found in nearly any environment.

Eukaryotes are a phylogenetic lineage characterized by the presence of a “true nucleus” surrounded by a double membrane. Early in eukaryote evolution, mitochondria were acquired by engulfing an alpha-proteobacterium and keeping it as a permanent endosymbiont (Figure I.1). No pre-mitochondrial extant eukaryotes are known, although several lineages have independently lost or reduced this organelle (Roger 1999; Baldauf et al. 2004). Photosynthetic eukaryotes (algae) are the result of a second fusion, involving a eukaryotic cell and a cyanobacterium (or a photosynthetic eukaryote, see below). These two endosymbiotic events shaped the complex evolution of photosynthetic eukaryotes.

### **Origin of photosynthetic eukaryotes**

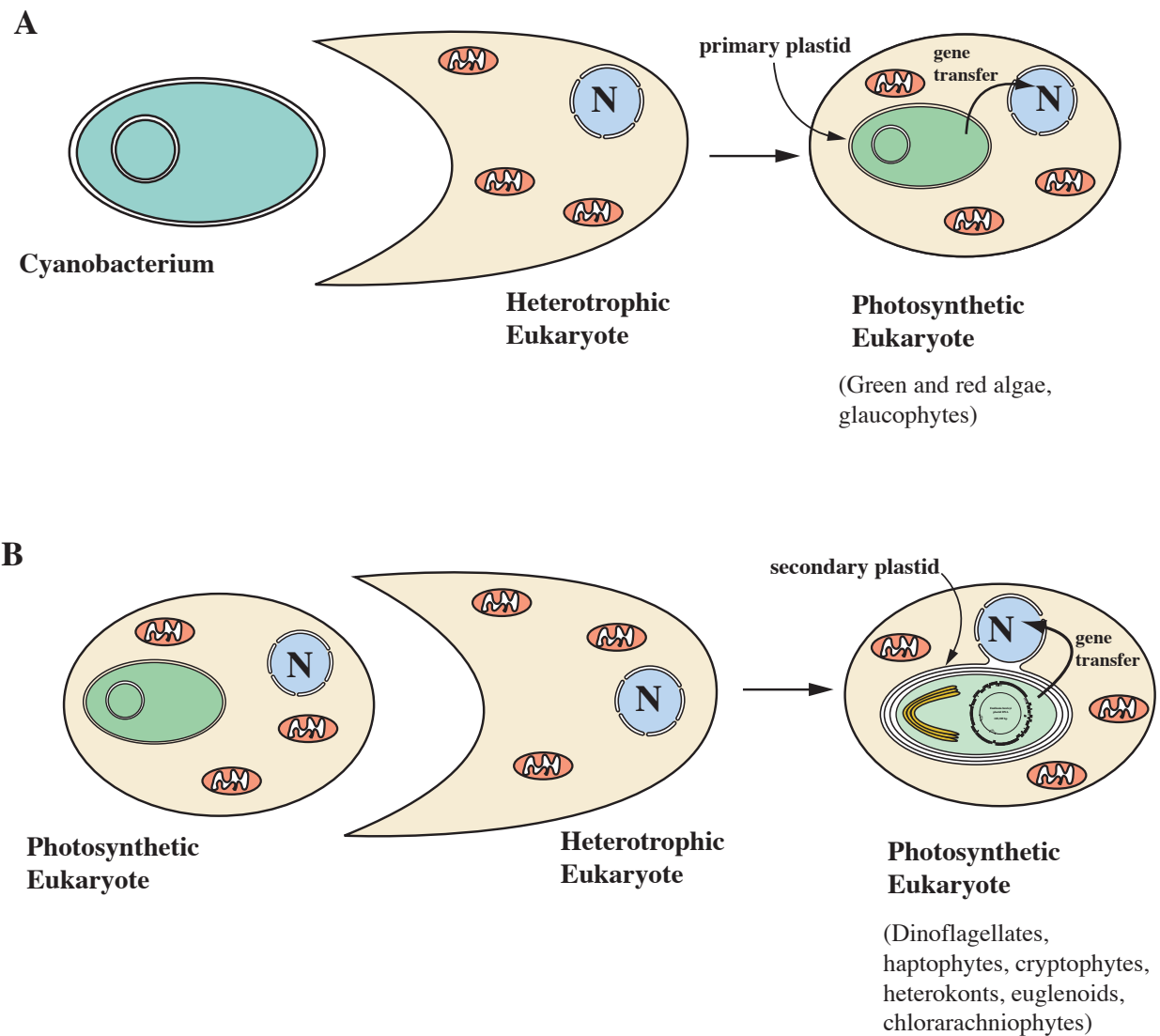
Photosynthesis is one of the most important processes that sustain life on earth. Different types of photosynthesis involving a variety of electron donors have been described in Bacteria and Archaea. Cyanobacteria are the only organisms capable of oxygenic photosynthesis, using water as electron donor by coupling photosystems I and II with solar radiation as a source of energy (Buchanan et al. 2000; Falkowski 2006). All photosynthetic eukaryotes have directly or indirectly acquired photosynthesis from a cyanobacterium and thus, are capable of oxygenic photosynthesis.

Photosynthetic eukaryotes have acquired plastids through the processes of primary, secondary, or tertiary endosymbiosis. Primary endosymbiosis is the process

by which a cyanobacterium was engulfed and integrated into a non-photosynthetic eukaryotic host cell, whereas in secondary endosymbiosis normally a non-photosynthetic eukaryotic host cell engulfed a photosynthetic eukaryote (Figure I.2). As a result, primary plastids, e.g. those of green and red algae, are surrounded by a double membrane derived from cyanobacterial membranes (Table I.1). The outer membrane, corresponding to the host phagosomal membrane, is lost (Bhattacharya et al. 2003). Secondary plastids, e.g. those of haptophyte, heterokonts, dinoflagellates, and cryptophytes, contain additional membranes derived from the endosymbiont plasma membrane and host endomembrane system. Following each endosymbiotic event, genes were transferred from the endosymbiont to the host nucleus. Therefore, plastids were free-living cyanobacteria, which subsequently lost the genes required for free-living existence.

Although part of the original cyanobacterial genome is maintained as a plastid genome, most cyanobacterial genes were exported to the nucleus (Palmer and Delwiche 1996; Martin et al. 1998; Martin et al. 2002; Brown 2003). Two hypotheses have been postulated regarding the mechanism of gene transfer to the host nuclear genome. The “bulk hypothesis” suggests that plastid DNA can be inserted in the nuclear genome following lysis of the organelle. An alternative hypothesis (“cDNA intermediate hypothesis”), holds that mRNA from plastid genes are involved in the process. Support for both hypotheses has been found, suggesting that both processes play a role in plastid-to-nucleus gene transfer (Nugent and Palmer 1991; Brown 2003; Timmis et al. 2004). Whichever mechanism served to transfer genes, protein products of nuclear-encoded genes are targeted back to the plastid and imported into the

Figure I.2. Acquisition of plastids in photosynthetic eukaryotes. A. Primary endosymbiosis. B. Secondary endosymbiosis. N= nucleus.

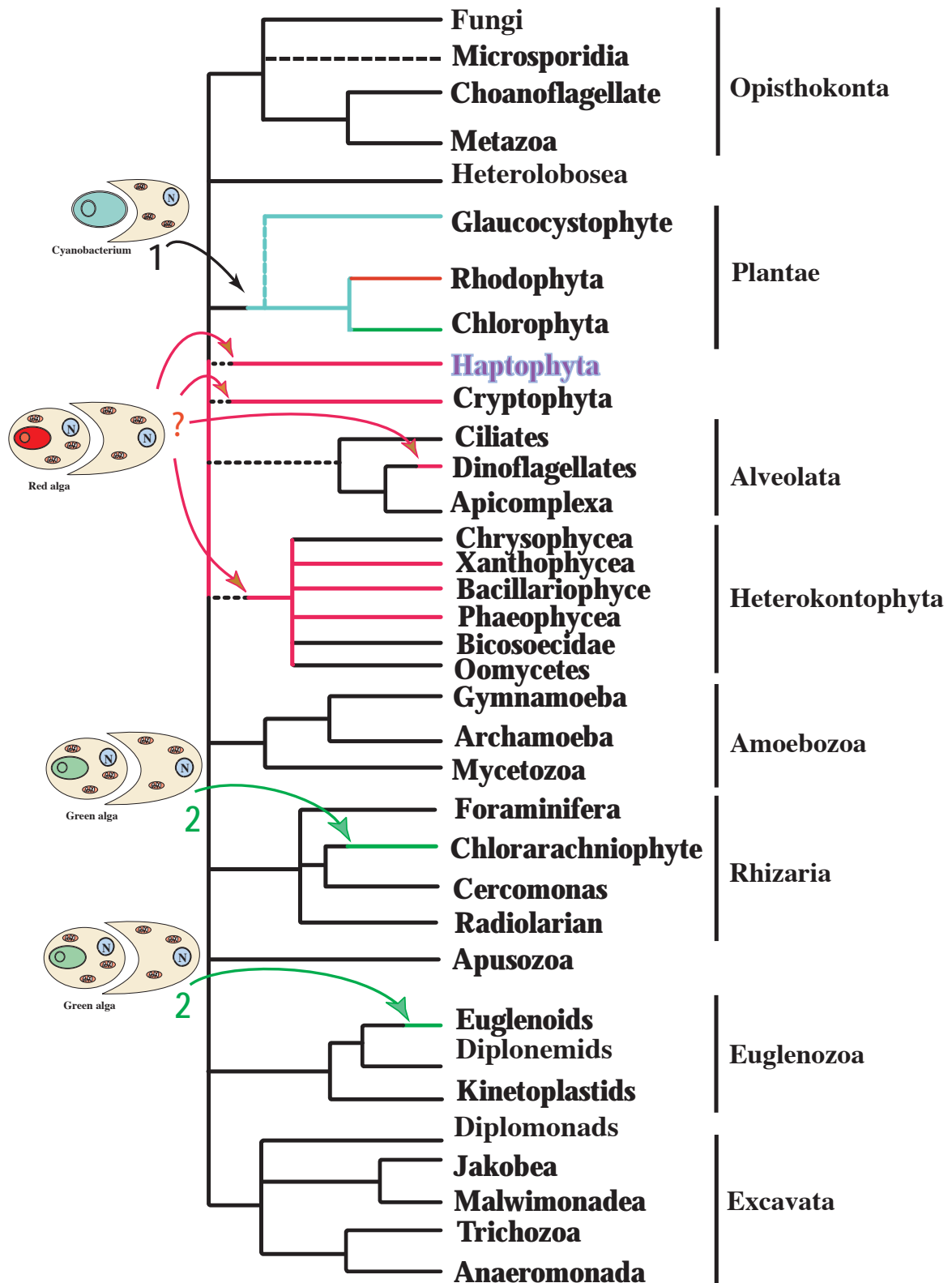




organelle aided by target localization signals and protein-import machinery (Martin et al. 1998; Cavalier-Smith 1999; Cavalier-Smith 2002). Targeting to primary plastids, such as those of green and red algae, requires a transit peptide to transport the proteins to the plastid and across the double membrane. In the case of secondary plastids, targeting signals consist of two parts, a signal and a transit peptide adjacent to one another, directing the protein first to the endomembrane system, and then to the plastid (van Dooren et al. 2001; Kroth 2002). Therefore, photosynthetic eukaryotes have fully integrated host and endosymbiont genomes. However, plastid and nuclear genome had previously independent evolutionary histories.

Evolutionary relationships among photosynthetic eukaryotes have often been confounded with the relationships among the plastids they harbor (Christensen 1962; Delwiche 1999; Palmer 2003). Once the endosymbiotic origin of plastids was described (Gibbs 1978; Gibbs 1981a), relationships of eukaryotic host cells came into question. Algae are regarded today as a polyphyletic clade with respect to the nuclear genome, but current data are insufficient to explain all relationships among photosynthetic eukaryotes. Phylogenetic analyses based on mitochondrial or nuclear sequence data have not resolved deep relationships among eukaryotes (Patterson 1999; Baldauf 2003; Cavalier-Smith 2003; Baldauf et al. 2004; Keeling 2004; Simpson and Roger 2004). However, major eukaryotic clades, defined in some cases primarily on molecular data, are recognized today, including Opisthokonta, Amoebozoa, Euglenozoa, Plantae, Heterokontophyta, and Alveolata, among others. A number of supergroups (Rhizaria, Excavata, and Chromalveolata) have also been proposed (Figure I.3), although the support for them is weak. More detailed studies

Figure I.3. Diagram of relationships among main eukaryotic lineages. Events of plastid acquisition through primary (1) or secondary (2) endosymbioses are shown. Question marks (?) indicate that it remains unknown the nature and number of endosymbiotic events that took place in those groups.



are necessary to assess whether these supergroups are natural groups or not and how they relate to each other.

Some eukaryotic lineages became photosynthetic by engulfing a cyanobacterium or a photosynthetic eukaryote, as discussed above (Figure I.3). A single primary endosymbiotic event probably gave rise to the three primary-plastid containing lineages, collectively called Plantae: green algae (including land plants), red algae and probably glaucophytes (Moreira et al. 2000; Rodriguez-Ezpeleta et al. 2005; Weber et al. 2006). Recently, a potential independent primary endosymbiosis has been described involving a cyanobacterium and the filose amoeba *Paulinella chromatophora* (Marin et al. 2005). At least three secondary endosymbiotic events are recognized today; two involving the engulfment of a green alga by an ancestral euglenoid and chlorarachniophyte in two separate events, and a minimum of one endosymbiosis involving the engulfment of a red alga, which gave rise to four eukaryotic lineages known as chromalveolates (Cavalier-Smith 1999; Delwiche 1999; Bhattacharya et al. 2003; Keeling 2004). Whether chromalveolates are monophyletic, and if so, the number of endosymbiotic events that took place in their evolution, along with the relationships among the four chromalveolate lineages, remain to be elucidated. These constitute the main questions driving my dissertation project.

## **Chromalveolates**

During the last decades, chromalveolate lineages have been grouped together, including or excluding some of their members, with the consequent erection of higher taxon names referring to overlapping groupings of taxa. Kingdom Chromista sensu

Table I.1. Description of plastids in several lineages of eukaryotes.

	Green algae	Red algae	Glaucophytes	Chlorarachnio phytes	Euglenoids	Cryptophytes	Haptophytes	Heterokonts	Dinoflagellates (peridinin-type)
Number of membranes	2	2	2 + peptido- glycan wall	4	3	4	4	4	3
Thylakoids per lamellae	many	single	single	1 to 3	3	2	3	3	3
Girdle lamella	no	no	no	no	no	no	no	yes	no
CER	no	no	no	no	no	yes	yes	yes	no
Nucleomorph	no	no	no	yes	no	yes	no	no	no
Chlorophyll types	a, b	a	a	a, b	a, b	a, c2	a, c1, c2, c3	a, c1, c2, c3	a, c2
Accessory pigments	carotenes, xanthophylls	phycobilisomes, carotenes, xanthophylls	phycobilisomes, carotenes, xanthophylls	xanthophylls	carotenoids, xanthophylls	phycobili- proteins, carotenes, xanthophylls	carotenes, fucoxanthin	fucoxanthin, vaucheriaxanthin	peridinin, carotenes
Storage	starch in chloroplast	floridean starch in cytosol	starch in cytosol	beta 1, 3 glucan in cytosol	paramylon in cytosol	starch in periplastid	chrysolaminarin or paramylon in cytosol	chrysolaminarin in cytosol	starch in cytosol
Plastid origin	cyanobacteria	cyanobacteria	cyanobacteria	green algae	green algae	red algae	red algae?	red algae?	red algae?
Plastid type	primary	primary	primary	secondary	secondary	secondary	secondary?	secondary?	secondary?

Cavalier-Smith (1989) includes cryptophytes, haptophytes, and heterokonts, based on the presence of tubular mastigonemes, the localization of the plastid in the endomembrane system of the host cell, or both. Mastigonemes are stiff tubular hairs with terminal hair-like extensions found on flagella of cryptophytes and heterokonts. Chl c containing algae comprise all photosynthetic eukaryotes with chl c, i.e. cryptophytes, heterokonts, haptophytes, and dinoflagellates (Bachvaroff et al. 2005). Chromalveolates (Chromista and Alveolata) is a more inclusive group that includes all chl c containing algae and their heterotrophic relatives (Cavalier-Smith 1999; Cavalier-Smith 2004), namely Haptophyta, Cryptophyta, Heterokontophyta (or Stramenopiles), and Alveolata (Dinophyta, Ciliophora, Apicomplexa).

Chromalveolates play critical ecological roles, contributing substantially to the primary production of the oceans and health of reef ecosystems, as well as being important grazers and parasites. Despite their environmental and ecological importance, little is known about the evolution of these organisms and the relationships among them. Historically, these four lineages have been grouped together based on their common pigmentation (Christensen 1962; Christensen 1989). However, when ultrastructural and molecular studies revealed that plastids were endosymbiotic cyanobacteria that could potentially be acquired in independent events, the value of plastid characters as phylogenetically informative features (for the organism) came into question. Today, the evolutionary relationships among cryptophytes, haptophytes, heterokonts, and alveolates are still controversial.

Plastids from chromalveolates are surrounded by three or four membranes, depending on the group, instead of two, as is the case of the primary plastids

contained in rhodophytes, chlorophytes, and glaucophytes (Table I.1) (Gibbs 1970; Delwiche 1999; Bhattacharya et al. 2003). When the four membranes are present, e.g. cryptophytes, haptophytes, and heterokonts, the inner two correspond to the dual membranes of the primary plastid (Figure I.4). The third membrane (periplastidal membrane) is thought to represent the plasma membrane of the endosymbiont and the fourth one is part of the endomembrane system of the host (Cavalier-Smith 1986; Palmer and Delwiche 1998). Evolutionary relationships of chl c containing plastids have been fairly well studied; a number of phylogenetic analyses based on plastid genes, including work derived from this dissertation, support the monophyly of the chl c containing plastids (Yoon et al. 2002b; Bachvaroff et al. 2005). However, most phylogenetic analyses have been based on a few genes and several taxa, or a large number of genes with only a reduced number of taxa, and conflicting topologies have been recovered (Martin et al. 1998; Fast et al. 2001; Ishida and Green 2002; Martin et al. 2002; Yoon et al. 2002b; Bachvaroff et al. 2005; Yoon et al. 2005). When this study began, no data from haptophyte organelles were available. In addition, the branching order within the chl c plastid clade has not been elucidated, and more data are required to test the different possibilities. To help orient the reader, a description of photosynthetic lineages of Chromalveolata follows.

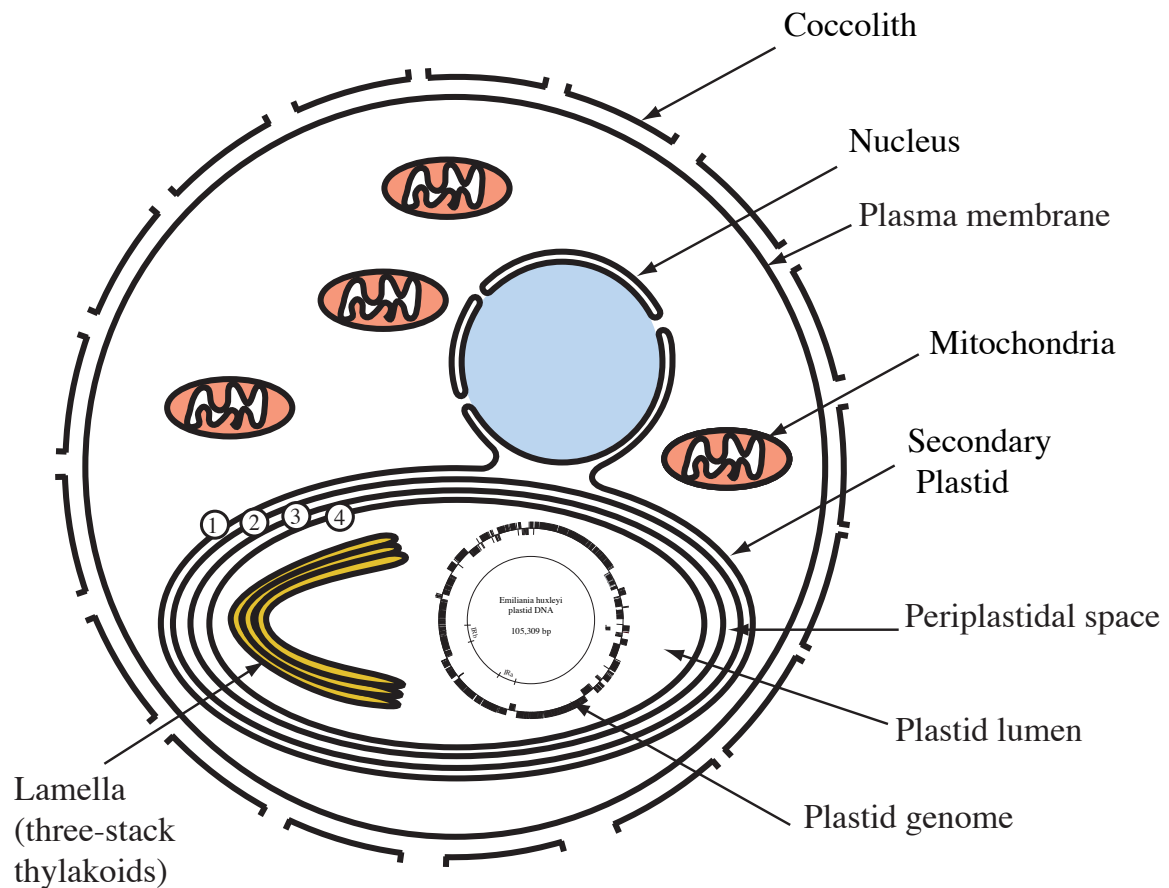
## **Haptophytes**

Most haptophytes are unicellular, photosynthetic eukaryotes, found mainly in marine environments, but also in freshwater. There are no well-documented heterotrophic members of this group (Marchant and Thomsen 1994), although many

of them are mixotrophic (Graham and Wilcox 2000; Andersen 2004). Representatives include motile (with two equal or subequal flagella) and non-motile forms. The Division Haptophyta includes more than 300 species and is divided into two major classes, Prymnesiophyceae and Pavlovophyceae, distinguished by the morphology of the flagellar apparatus and mitotic spindle (Green and Jordan 1994). Class Prymnesiophyceae can be further divided in four orders: Phaeocystales, Prymnesiales, Coccolithales, and Isochrysidales (Edwardsen et al. 2000). Other clades have been recognized through phylogenetic analyses, but new orders have not been formally proposed (Saez et al. 2004). The coccolithophorids (“calcium carbonate scale-bearing cells”), the main group within the Prymnesiophyceae, are important contributors to the vertical flux of carbon in the ocean and are capable of forming large blooms in all oceans. In addition, they are, in part, responsible for the production of dimethylsulphide (DMS), which is involved in cloud nucleation (Malin and Kirst 1997; Falkowski et al. 2000; Riebesell et al. 2000). They are regarded as the algal group with the single most important impact on long-term carbon and sulfur cycling (Graham and Wilcox 2000).

Ultrastructural and molecular evidence indicate that the haptophytes are a monophyletic group with the primary morphological synapomorphy being a characteristic appendage, the haptonema (Christensen 1962; Christensen 1989; Green and Jordan 1994). The haptonema is composed of 6-7 microtubules arranged as a crescent or a ring, surrounded by the cell membrane, with the length depending on the species. A number of functions have been attributed to this appendage, including food capture or attachment to the substratum (Inouye and Kawachi 1994). Haptophyte

Figure I.4. Diagram of a cell of the cocolithophorid *Emiliana huxleyi*. Four membranes surround the plastid: 1- Chloroplast ER (CER), derived from the endomembrane system of the host cell (haptophyte), continuous with the nuclear envelope and the endoplasmic reticulum (ER); 2- remnant of the endosymbiont plasma membrane, named periplastid membrane (PPM); 3 and 4- membranes from the primary plastid that were derived from cyanobacterial membranes.





plastids are pigmented with chlorophyll a and c (Table I.1), and two related carotenoid fingerprints, 19'hexanoyloxyfucoxanthin and 19' butanoyloxyfucoxanthin (Bijornland and Liaaen-Jensen 1989). The plastids are surrounded by four membranes (Figure I.4), from which the outermost membrane is continuous with the ER and is called the chloroplast ER (CER) (Gibbs 1981b). Thylakoids are stacked in groups of three to form lamellae; no girdle lamella is present (Table I.1). Coccolithophorids are a monophyletic group, most of whom are covered by several layers of calcium carbonate scales (coccoliths), although a few species have secondarily lost these scales, and not all species produce them in all phases of the life history. Coccoliths are readily preserved with the earliest fossil record from ca. 100 Ma (Ziveri et al. 2004). Because coccoliths are characteristic of a derived group of haptophytes, this date provides a minimum age for coccolithophorids, but is probably a significant underestimate for the age of haptophytes as a lineage.

Little is known about the evolution of the haptophytes and their phylogenetic relationships to other living organisms. Studies using nuclear genes support the distinctiveness of the haptophytes, but have not shown their affinity to any other group (Bhattacharya et al. 1993; Medlin et al. 1996; Medlin et al. 1997; Tengs et al. 2000; Ali et al. 2001; Stechmann and Cavalier-Smith 2003). Despite the importance of these algae, as of 2002, no organellar or nuclear genome from any member of this group had been sequenced. The lack of molecular data from haptophytes accounts for some of the unresolved questions in chromalveolate evolution. Chapters 2 and 3 of this dissertation report the mitochondrial and plastid genomes of the haptophyte

*Emiliana huxleyi*, respectively. The complete nuclear genome of *E. huxleyi* is presently being sequenced elsewhere.

## **Dinoflagellates**

Dinoflagellates are biflagellate protists that live in freshwater and marine environments; most species are unicellular although some are coccoid, filamentous, or coenocytic in some stages of their life history. These organisms are mostly known for their characteristic blooms, named red tides, which can affect seafood production and may be toxic. Around 3,000 extant species of dinoflagellates have been described and more than 2,000 fossil species are known, belonging to about 130 genera (Taylor 1987). The earliest undisputed dinoflagellate fossils are from 400 Ma, but the major radiation occurred ca. 200 Ma. (Fensome et al. 1999). Most dinoflagellate cells contain two grooves: a cingulum (equatorial groove), and a sulcus (longitudinal groove on the ventral side). Two distinct flagella are present in the cell; a longitudinal flagellum points backwards and runs along the sulcus, and a transversal coiled flagellum lies in the cingulum (Graham and Wilcox 2000). Dinoflagellate cells contain one layer of vesicles (alveoli) below the plasma membrane, which in some cases are filled with cellulose to form a theca (“armored dinoflagellates”), but may be empty or filled with such thin layers of cellulose that they are not visible under the light microscope (“naked dinoflagellates”). In either case, the cortical structure is thought to give shape to the cell. Taxonomy of thecate dinoflagellates has been based mainly on the pattern and number of thecal plates (Taylor 1987). “Naked”

dinoflagellates can also be identified with this system, albeit with somewhat more difficulty.

The nucleus of dinoflagellates has many peculiar features, including permanently condensed chromosomes, a large quantity of DNA, absence of nucleosome structures and most histones, and nuclear DNA associated with bacterial histone-like proteins (Rizzo 1987; Wong et al. 2003). Some histones are also thought to be present, but exactly how the DNA is organized remains poorly understood (Okamoto and Hastings 2003; Hackett et al. 2005). Dinoflagellates have diverse ecological roles and nutritional strategies: photosynthetic, mixotrophic, predatory, and parasitic. About half of the known species are nonphotosynthetic and the remainder have some sort of plastid and rely entirely or partially on photosynthesis. Only a few photosynthetic species are obligate autotrophs and the majority are mixotrophs (Schnepf and Elbrächter 1999). Photosynthetic dinoflagellates form a monophyletic clade sister to several heterotrophic lineages (Gunderson et al. 1999; Leander and Keeling 2004). No genomic data are available from heterotrophic dinoflagellates; a few EST (expressed sequence tag) projects have recently been conducted on photosynthetic ones (Bachvaroff et al. 2004; Tanikawa et al. 2004; Hackett et al. 2005; Lidie et al. 2005). Chapter V reports analysis of an EST project from the non-photosynthetic dinoflagellate *Cryptocodinium cohnii*.

Most photosynthetic dinoflagellates have plastids that are surrounded by three membranes and contain chlorophylls *a*, *c*, and peridinin as the major photosynthetic pigments (Table I.1). The outermost membrane in dinoflagellates lacks ribosomes, and it is not known whether it corresponds to the periplastid or the phagosomal

membrane (Cavalier-Smith 1999; Bhattacharya et al. 2003). The organization of plastid genes in peridinin-containing dinoflagellates has been characterized in *Heterocapsa triquetra*, *Amphidinium operculatum*, and *A. carterae*. They apparently lack a conventional plastid genome and have instead several small circular DNA molecules, typically about 2-3 kbp, containing 0, 1, or 2 genes, with fewer than 20 plastid genes identified so far (Zhang et al. 1999; Barbrook and Howe 2000; Hiller 2001). These genes have been presumed to be located in the plastid (Takishita et al. 2003), but one study suggests nuclear localization (Laatsch et al. 2004). Recently, a study of the dinoflagellate *Gonyaulax polyedra* revealed that the plastid gene *psbA* is not encoded on a minicircle, and it is associated with DNA of 50-150 kb (Wang and Morse 2006). Peridinin-containing plastids are also peculiar in the use of a nuclear-encoded form II rubisco similar to that of alpha-proteobacteria instead of a plastid-encoded form Ia rubisco, as in all other red-algal derived plastids (Morse et al. 1995; Delwiche and Palmer 1996). The phylogenetic origin of the peridinin-containing dinoflagellate plastid remains unknown due partly to the extreme rate of evolution of minicircle genes (Zhang et al. 2000; Yoon et al. 2002a; Bachvaroff et al. 2006).

A small fraction of the photosynthetic species of dinoflagellates have plastids not containing peridinin, but with a pigment composition characteristic of another algal group (Dodge 1975; Chesnick et al. 1997; Schnepf and Elbrächter 1999; Takishita et al. 2000; Tengs et al. 2000). These anomalously pigmented plastids have been acquired from other algal groups, including cryptophytes, heterokonts, haptophytes, and green algae by means of secondary or tertiary endosymbioses. Haptophyte-containing dinoflagellates have 19'-hexanoyloxyfucoxanthin and 19'

butanoyloxyfucoxanthin as accessory pigments and comprise members of at least three genera that form a monophyletic group, suggesting a single acquisition of the fucoxanthin-containing plastids (Tengs et al. 2000).

Apicomplexa are the closest relatives to dinoflagellates, and together with ciliates are collectively called Alveolates, due to the presence of cortical alveoli in these three lineages (Cavalier-Smith 1991). Ciliates are heterotrophic and aplastidic, while apicomplexans are parasitic and contain a reduced plastid, called an apicoplast, surrounded by four membranes (Köhler et al. 1997; Foth and McFadden 2003). Due to the medical importance of apicomplexan parasites, many genetic and genomic studies of these organisms have been performed, including the complete sequence of three apicoplast, five mitochondrial, and more than five nuclear genomes from members of this group. In contrast, only three mitochondrial genomes from ciliates have been sequenced and the complete nuclear genome of *Tetrahymena pyriformis* is under way. The origin of the apicoplast and its relationship to dinoflagellate plastids is not clear. Some evidence supports a green algal origin, but most data suggest a red algal ancestry of the apicoplast (Köhler et al. 1997; Fast et al. 2001; Funes et al. 2004). A single acquisition of plastids in the common ancestor of apicomplexans and dinoflagellates is under dispute and awaits testing.

## **Cryptophytes**

Cryptophytes comprise a small group (ca. 200 species) of photosynthetic and heterotrophic organisms that live in marine and freshwater environments. They are unicellular flagellates, with two flagella emerging from an apical vestibulum. The

flagella bear two-parted flagellar hairs (mastigonemes). Cryptophyte cells are surrounded by a proteinaceous periplast inside the plasma membrane (Graham and Wilcox 2000; Adl et al. 2005). Except for the heterotrophic genus *Goniomonas*, cryptophytes contain secondary plastids derived from the red algae (McFadden et al. 1994b). Plastid-containing lineages of cryptophytes are monophyletic to the exclusion of basal heterotrophic taxa (McFadden et al. 1994b; Marin et al. 1998). Cryptophyte plastids contain chl a, c, and phycobiliproteins, which are not arranged in phycobilisomes, but rather localized in the thylakoid lumen (Table I.1). Plastids are surrounded by four membranes with the outermost membrane continuous with the ER (CER). Thylakoids are arranged in pairs to form lamellae; no girdle lamella is present. In the periplastidal space (between the inner two and the outer two membranes), cryptophytes still maintain a remnant of the nucleus of the red algal endosymbiont called a “nucleomorph” (Gibbs 1962; Guillot and Gibbs 1980a; Guillot and Gibbs 1980b). This highly reduced eukaryotic genome consists of three fully sequenced chromosomes that encode mostly “housekeeping” genes (Maier et al. 2000; Gilson 2001). Phylogenetic analyses based on these data support the hypothesis that cryptophytes acquired their plastids from the red algal lineage (Eschbach et al. 1991; Cavalier-Smith et al. 1994; McFadden et al. 1994a; Gilson and McFadden 1996). Along with the complete sequence of the nucleomorph, one plastid and one mitochondrial genome sequence are available for this group.

## Heterokonts or Stramenopiles

Phylum Heterokontophyta (“algae with different flagella”) encompasses an extremely diverse group of algae that inhabit both marine and freshwater environments, ranging from unicellular diatoms (Class Bacillariophyceae) to giant kelps (Class Phaeophyceae). It comprises around 100,000 species classified in twelve classes (Patterson 1999; Graham and Wilcox 2000; Adl et al. 2005), with photosynthetic and heterotrophic members, as well as parasitic ones (e.g. *Phytophthora infestans*, causative agent of the potato late blight). Basal heterokonts, including bicosoecids, labyrinthulids, and oomycetes are heterotrophic (Cavalier-Smith and Chao 1996; Karpov et al. 2001). Stramenopiles are characterized by flagellate stages with two distinct flagella (although some derived members of the group have other flagellar organizations). A long forward directed flagellum bears two rows of tripartite tubular hairs (mastigonemes), and a smooth posterior flagellum exhibits a swelling that is often associated with a light-sensing system (van den Hoek et al. 1995). Plastids are located within the ER and are surrounded by four membranes (Table I.1). They contain chl a, c, and a diverse range of accessory pigments depending on the group. The major pigment is fucoxanthin in diatoms, chrysophytes, and phaeophytes, while vaucheriaxanthin is most common in raphidophytes, eustigmatophytes, and tribophytes. In all groups, thylakoids are stacked in groups of three, and a girdle lamella runs beneath the innermost plastid membrane (van den Hoek et al. 1995). Stramenopile fossils are abundant from the past 150 Ma, although recent discoveries suggest this lineage has existed for ca 500 Ma and fossils from 1,000 Ma resemble extant species of heterokonts (Xiao et al. 1998). Genomic data

for this large group of algae are not abundant but there is relatively more information than for any of the other chl c containing lineages. Three heterokont nuclear genomes are being fully sequenced (centric diatom *Thalassiosira pseudonana*, pennate diatom *Phaeodactylum tricornutum*, and oomycete *Phytophthora infestans*), and one plastid and 11 mitochondrial complete genomes are available.

### **Hypotheses of plastid evolution in chromalveolates**

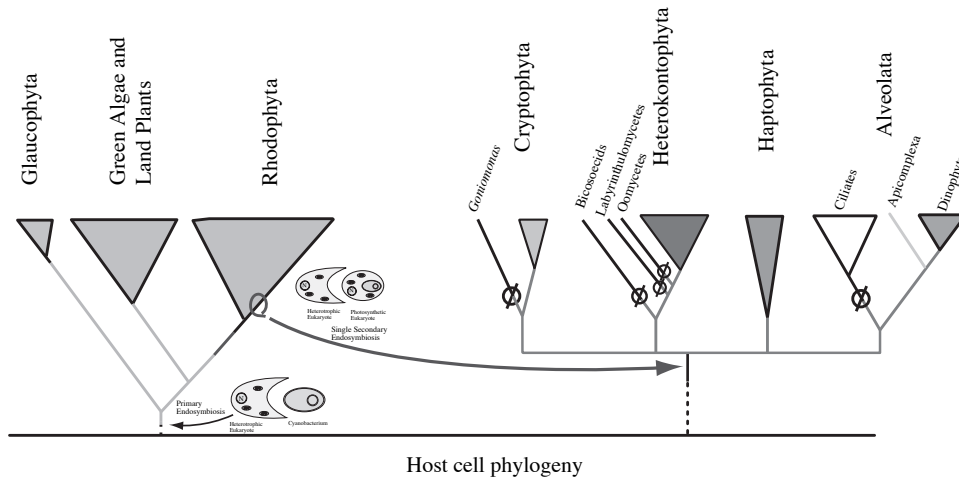
Understanding the evolution of chromalveolates, including the events of plastid acquisition, requires knowledge of the evolutionary relationships among their plastids, as well as those of the host cell lineages. Good evidence, including that presented in this dissertation, exist for a single origin of chl c plastids, but relationships within the chl c plastid clade remain poorly understood, as well as the relationships among the host cells. Knowledge of chl c plastid relationships conveys only partial information regarding the number of endosymbiotic events that took place in chromalveolate evolution, or the relationships among their host genomes. Several models of host cell evolution can be postulated, assuming a monophyletic chl c plastid clade (Figure I.5). The reduced genetic information available from haptophytes, as well as from dinoflagellates, is one of the main reasons why this is still controversial. Interestingly, cryptophytes, heterokonts, and dinoflagellates have basal members that are known to be heterotrophic (Cavalier-Smith and Chao 1996; Van de Peer and De Wachter 1997; Marin et al. 1998; Gunderson et al. 1999).

There are three competing hypotheses regarding the number of endosymbiotic events in chromalveolates and the relationships among them (Figure I.5). In the first

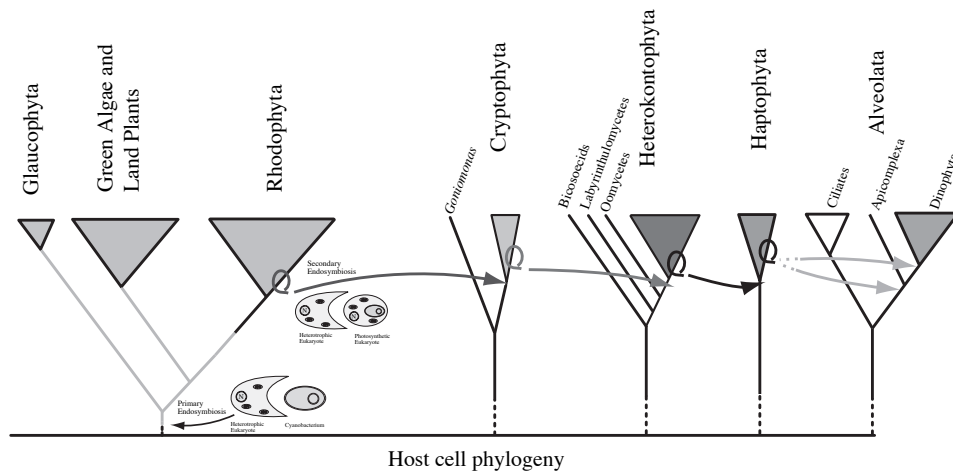


Figure I.5. Proposed models of chromalveolate evolution, congruent with the monophyly of the chl *c* containing plastids. A- Chromalveolate hypothesis. B- Serial plastid transfer hypothesis. C- Parallel plastid transfer hypothesis. Putative plastid losses are indicated by crossed circles.

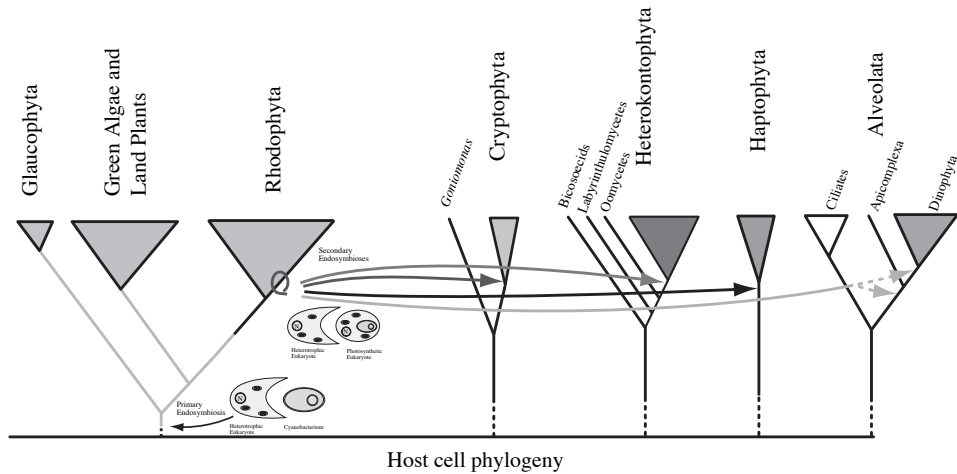
A



B



C



of these (chromalveolate hypothesis, Figure I.5.A), a single endosymbiotic event for all chromalveolates occurred between a unicellular red alga and a biciliate protozoan before the divergence of the four chromalveolate lineages. This would imply that the host cells (chromalveolates) are monophyletic, and that plastids were lost many times in chromalveolate evolution (Cavalier-Smith 1981; Cavalier-Smith 1989; Cavalier-Smith 1999; Cavalier-Smith 2002). An alternative hypothesis (serial plastid transfer hypothesis, Figure I.5.B) infers multiple independent endosymbiotic events in chromalveolate lineages, allowing host and endosymbiont phylogenies to be incongruent (Cavalier-Smith et al. 1994; Palmer and Delwiche 1998). This hypothesis implies a non-photosynthetic ancestor for each chromalveolate lineage, and thus, basal heterotrophic members never contained a plastid. Under this model, one chromalveolate lineage, e.g. cryptophytes, acquired the plastid from a red alga in a secondary endosymbiosis. Later, another chromalveolate lineage engulfed the newly photosynthetic chromalveolate in a tertiary endosymbiotic event, and so on with the remaining lineages. A third hypothesis (parallel plastid transfer hypothesis, Figure I.5.C) postulates independent plastid acquisitions from closely related red algae by all four chromalveolate lineages. Under this model, chromalveolates may or may not be monophyletic and basal heterotrophic lineages may have never contained a plastid. This hypothesis is the least likely because it invokes convergent evolution of chlorophyll c biosynthesis and plastid targeting mechanism. In addition, recent studies based on the gene GAPDH (glyceraldehyde-3P-dehydrogenase) argue against this model (Fast et al. 2001; Harper and Keeling 2003).

Most authors seem to agree that the question remains unresolved, with some authors expressing very strong contradictory opinions, often on the bases of relatively sparse data. The complexity of the process of plastid acquisition and organelle formation (i.e. transfer of hundreds of plastid genes to the nucleus and acquisition of signal sequences) argues for a reduced number of independent plastid acquisitions (Cavalier-Smith et al. 1994; Cavalier-Smith 1999). However, by minimizing the endosymbiotic events, the number of plastid losses required to explain the phylogeny is necessarily increased. To evaluate these hypotheses, it is important to understand how readily plastids can be acquired or lost in evolution. Likewise, understanding plastid and host cell phylogenetic relationships will allow us to assess the relative importance and complexity of plastid acquisition and plastid loss in evolution.

### ***Objectives and Significance***

The main objectives are:

1. Acquisition of organellar sequence data from a representative of Haptophyta, description of fundamental genomic features, and comparison of genetic characteristics, such as gene content and gene clusters, with other eukaryotic lineages.
2. Investigation of phylogenetic relationships among photosynthetic eukaryotes to test hypotheses of eukaryotic evolution.
3. Analysis of evolutionary relationships of chromalveolate plastids and interpretation of pattern of plastid acquisition in chromalveolates.

4. Examination of nuclear data from a heterotrophic dinoflagellate and evaluation of evidence for previous plastid endosymbiosis.
5. Elucidation of plastid-related metabolic pathways in a heterotrophic dinoflagellate.

The initial objective of my project was to examine the fundamental properties of the organellar genomes from haptophytes and explore its suitability for genomic studies (Chapters 2 and 3 of this dissertation, and published as Sanchez Puerta et al. 2004, 2005, as well as some material presented in Bachvaroff et al. 2005). Because relatively little genetic work has been performed with haptophytes, the project provides basic background information on the gene content and organization of plastid and mitochondrial genomes, as well as fundamental genetic properties of these organelles, such as base composition and codon usage. The data presented here constitute the first complete organellar genomes from the phylum and represent a unique opportunity to increase our understanding of the biology and evolution of the Haptophyta. At the same time, these newly acquired sequence data allow us to perform more comprehensive phylogenetic analyses including all four chromalveolate lineages.

One question I want to address is the phylogenetic position of haptophytes and their relationships with other eukaryotic lineages (Chapter II). This is an ambitious task because haptophytes are a very distinctive group and have been shown to be particularly challenging to place, probably due to the rapid radiation of several groups of eukaryotes in a short period of time. Furthermore, acquisition of genomic data

from haptophytes represents a key piece needed to solve the puzzle of chromoalveolate evolution. Mitochondrial genome analysis is a valuable tool for resolving evolutionary relationships among various eukaryotic lineages (Gray et al. 1998; Gray et al. 1999). In this work, I perform phylogenetic analyses based on a number of mitochondrial genes to assess the evolutionary relationships of the haptophyte host cell.

Another issue I focus on is the origin of the haptophyte plastids and their relationships to the other chlorophyll c containing plastids (Chapter IV). The most direct way to investigate this is to acquire a significant number of plastid gene sequences from all chl c containing lineages. As of 2002, there were only a few haptophyte plastid sequences available while organellar genomes of representatives of the red algae, cryptophytes, and heterokonts had been fully sequenced. In addition, plastid gene data and two EST projects in dinoflagellates became available. Therefore, the complete sequence of the plastid genome of a member of the haptophytes is of great importance and constitutes an invaluable source of information to understand the origin and relationships of the chl c plastids.

Last but not least, I analyzed an expressed sequence tag (EST) project of the heterotrophic dinoflagellate *Cryptothecodinium cohnii* (Chapter V). This study has intrinsic importance for being the first genomic study of a non-photosynthetic dinoflagellate. In addition, *C. cohnii* has been described as an early divergent species, sister to all photosynthetic dinoflagellates, and thus, the presence of plastid-associated genes in *C. cohnii* indicates an earlier plastid acquisition by this group than would otherwise be inferred. I focused on plastid-related genes to map plastid-associated

pathways remnant of a previous endosymbiosis, which are maintained in this heterotrophic species.

## Chapter II – The Complete Mitochondrial Genome

### Sequence of the Haptophyte *Emiliana huxleyi* and its

### Relation to Heterokonts

#### *Abstract*

The complete nucleotide sequence of the mitochondrial genome of *Emiliana huxleyi* (Haptophyta) was determined. *E. huxleyi* is the most abundant coccolithophorid, key in many marine ecosystems and plays a vital role in the global carbon cycle. The mitochondrial genome contains genes encoding three subunits of the cytochrome c oxidase, apocytochrome b, seven subunits of the NADH dehydrogenase complex, two ATPase subunits, two ribosomal RNAs, 25 tRNAs and five ribosomal proteins. One potentially functional open reading frame was identified, with no counterpart in any other organism so far studied. The *cox1* gene transcript is apparently spliced from two distant segments in the genome. One of the most interesting features in this mtDNA is the presence of the *dam* gene, which codes for a DNA adenine methyltransferase. This enzyme is common in bacterial and archaeal genomes, but is not present in any studied mitochondrial genome. Despite the great age of this group (ca. 300 Ma), little is known about the evolution of haptophytes or their relationship to other eukaryotes. This is the first published haptophyte organellar genome, and will improve the understanding of their biology and evolution and allow us to test the monophyly of the chromoalveolate clade.

## ***Introduction***

*Emiliania huxleyi* (Lohmann) Hay & Mohler is the most abundant of the coccolithophorids, a key group of marine phytoplankton. It has been the subject of numerous studies, but only a few of these employed genetic approaches. This species has been relatively well studied because of its potential importance in the global carbon cycle (Falkowski et al. 2000; Riebesell et al. 2000). It is capable of forming large blooms in all oceans, particularly at mid latitudes, can reach cell densities of  $10^7$  cells/L and cover thousands of square kilometers (Balch et al. 1992; Brown and Yoder 1994; Winter and Siesser 1994; Green and Harris 1996; Paasche 2002). *E. huxleyi* blooms emit large amounts of dimethylsulfide (DMS), which upon oxidation in the atmosphere is considered an active component in the nucleation of refractive clouds (Malin et al. 1992; Malin and Kirst 1997). Therefore, this species is regarded as a key component of the greenhouse effect, natural acid rain, and albedo regulation.

Little is known about the evolution of haptophytes and their phylogenetic relationships to other living organisms (Bhattacharya et al. 1993; Medlin et al. 1996; Medlin et al. 1997; Tengs et al. 2000). On the basis of pigmentation, it has been suggested that the haptophytes belong to a monophyletic group of organisms with chlorophyll c containing plastids, called chromoalveolates (Cavalier-Smith 1981; Cavalier-Smith 1989; Cavalier-Smith 1999), but the evolutionary history of plastids does not necessarily reflect that of the whole cell. Mitochondrial genome analysis has been recognized as a valuable tool for resolving evolutionary relationships among the various eukaryotic lineages (Gray et al. 1998; Gray et al. 1999). The diversity in mitochondrial genome size, gene content, and organization is an important tool to



elucidate the mechanisms and reconstruct the pathway by which this evolutionary diversification has occurred (Gray et al. 1998). Because this is the first genome to be sequenced in the phylum Haptophyta, it is a unique opportunity to increase our understanding in the biology and evolution of the members that comprise this group, in particular the widely distributed and environmentally important *E. huxleyi*. The phylogenetic position of haptophytes has been examined with single-gene phylogenies (Bhattacharya et al. 1993; Medlin et al. 1996; Medlin et al. 1997; Gray et al. 1998). The approach taken here is to understand protist evolution using comparative analyses of whole mitochondrial genomes, which allow comprehensive phylogenetic analysis. This chapter (published as Sanchez-Puerta et al. 2004) presents the complete sequence of the mitochondrial DNA (mtDNA) of the coccolithophorid *E. huxleyi*, description of the main features, and phylogenetic hypotheses derived from the newly available data.

## ***Materials and Methods***

The complete mtDNA sequence of *Emiliania huxleyi* has been deposited in GenBank (accession number AY342361).

## **Culture of *E. huxleyi* and mtDNA isolation**

The axenic strain of *E. huxleyi* was obtained from Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP # 373). Cultures were grown in Guillard's f/2 medium (Andersen et al. 1997) at 17°C with a 14h/10h L:D cycle. Approximately 6 L of culture were harvested by centrifugation, flash frozen in liquid

nitrogen, and stored at -80°C. For total DNA extraction and organellar DNA purification, the protocol designed by Chesnick and Cattolico (Chesnick and Cattolico 1993) was followed. Frozen algal tissue was ground and cells were lysed by adding 120 ml of 2% CTAB detergent and kept at 60°C for 20 minutes. Then, 80 ml of chloroform were added and mixed by inversion. I centrifuged the sample at 6,000 rpm for 15 minutes, and the DNA in the supernatant was precipitated by adding two volumes of 100% ethanol. The DNA was resuspended in 3 ml of TE, combined with CsCl and 10 ul of Hoechst 33258 dye (Molecular Probes, Inc, Oregon) to reach a density of 1.6 g/ml. Mitochondrial, chloroplast, and nuclear DNA were separated through CsCl-bisbenzimidazole isopycnic centrifugation in TLA-110 Rotor at 70 krpm for 17 hours, by which mtDNA forms the least dense band. Bands of DNA were collected by cutting the top of the centrifuge tube and aspirating the DNA with a needle and a syringe.

### **Cloning and DNA sequencing**

Mitochondrial DNA was digested with the restriction endonuclease *HindIII* and the resulting fragments were cloned in pGEM -3Zf(+) (Promega, WI) using *Escherichia coli* XL-10 Gold Ultracompetent Cells (Stratagene, CA) as the host bacterium. Plasmids from individual clones were isolated using the ‘miniprep’ procedure (Sambrook and Russell 2001), and sequenced using dye terminator chemistry (ABI). The M13-20 primer was used for 5’ and T7 primer for 3’ sequencing. Primer walking was used to determine the full sequence of longer clones and to

obtain double stranded sequencing reads. The polymerase chain reaction was used to order the fragments, fill gaps, and obtain double stranded coverage.

## **Data analysis**

Sequences were edited using the program Sequencher (GeneCodes Corp., Ann Arbor, MI). Vector and low quality bases were removed, and manual editing was performed. Sequence reads were assembled using the contig assembly function of Sequencher.

Putative open reading frames (ORFs) were identified by performing BLAST searches of the GenBank databases at the National Center for Biotechnology Information (NCBI). Similarity searches to detect tRNAs were performed with tRNAscan SE Search Server (Washington University, St. Louis; <http://www.genetics.wustl.edu/eddy/tRNAscan-SE/>). General codon usage analyses were performed using GCUA (McInerney 1998). Correspondence analysis of codon usage by genes and the indices, frequency of optimal codons (Fop) and codon adaptation index (CAI), were calculated using CodonW (University of Nottingham; <http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html>).

## **Phylogenetic trees**

Thirteen mitochondrial genes (*atp6*, *atp9*, *cob*, *cox1*, *cox2*, *cox3*, *nad1*, *nad2*, *nad3*, *nad4L*, *nad5*, *nad6*, *rps12*) were first aligned using individual protein alignments with ClustalW ([www.cmbi.kun.nl/bioinf/tools/clustalw.shtml](http://www.cmbi.kun.nl/bioinf/tools/clustalw.shtml)) output as a

starting point, then manually edited with MacClade 4.0 (Maddison and Maddison 2000). Unalignable regions were excluded and protein-coding genes were concatenated as a nucleotide alignment. Table II.1 lists the accession numbers in GenBank of the species included in the phylogenetic analyses. Individual, as well as concatenated, gene phylogenetic analyses have been performed. Two concatenated datasets were constructed: 4-gene dataset (*cob*, *cox1*, *cox2*, *cox3*; 1236 aa) and 13-gene dataset (all genes included, 2787 aa). Analyses were based on both nucleotide (with and without the third codon position) and amino acid data. For maximum likelihood (ML) nucleotide analysis using PAUP\* parameters were estimated from a Fitch-Margoliash tree using LogDet distances. In the likelihood analysis, the General Time Reversible model with Invariant site and gamma correction with four rate categories (GTR + I +  $\Gamma$ 4) was used. Bootstrap analyses were performed using three random additions with nearest neighbor interchange. Bayesian analysis was performed using the MrBayes program (Huelsenbeck and Ronquist 2001) with the same likelihood model as the PAUP\* searches, i.e., GTR+I+ $\Gamma$ 4. Four Markov chains were run with one heated for  $2 \times 10^6$  generations, sampled every hundredth generation with a burnin period of 1,000 generations. For ML protein analyses, the PhyML program (Guindon and Gascuel 2003) was used with four gamma distributed rates and invariant sites under the JTT model of amino acid substitution (Jones et al. 1992).

Table II.1. Sources of sequences used in phylogenetic analyses

Taxon Name	GenBank Accession Number
<i>Acanthamoeba castellani</i>	NC_001637
<i>Allomyces macrogynus</i>	NC_001715
<i>Amphidinium carterae</i>	CF065611, CF065139, CF065342, CF067399, CF064778, CF065944
<i>Beta vulgaris</i>	NC_002511
<i>Cafeteria roenbergensis</i>	NC_000946
<i>Chaetosphaeridium globosum</i>	NC_004118
<i>Chondrus crispus</i>	NC_001677
<i>Chrysodidymus synuroideus</i>	NC_002174
<i>Cyanidioschyzon merolae</i>	NC_000887
<i>Cyanidium caldarium</i>	Z48930
<i>Emiliana huxleyi</i>	AY342361
<i>Gracilariopsis lemaneiformis</i>	AF118119
<i>Homo sapiens</i>	NC_001807
<i>Laminaria digitata</i>	NC_004024
<i>Malawimonas jakobiformis</i>	NC_002553
<i>Monosiga brevicollis</i>	NC_004309
<i>Ochromonas danica</i>	NC_002571
<i>Paramecium aurelia</i>	NC_001324
<i>Phytophthora infestans</i>	NC_002387
<i>Plasmodium falciparum</i>	NC_002375
<i>Platymonas subcordiformis</i>	Z47795, Z47797
<i>Podospora anserina</i>	NC_001329
<i>Porphyra purpurea</i>	NC_002007
<i>Prototheca wickerhamii</i>	NC_001613
<i>Pylaiella littoralis</i>	NC_003055
<i>Reclinomonas americana</i>	NC_001823
<i>Rhodomonas salina</i>	NC_002572
<i>Tetrahymena thermophila</i>	NC_000862
<i>Toxoplasma gondii</i>	AAF04081, AAC34138, AAF07939, CB384139
<i>Triticum aestivum</i>	AF337547, Y00417, AF336134, X15944
<i>Xenopus laevis</i>	NC_001573
<i>Dictyostelium discoideum</i>	NC_000895
<i>Naegleria gruberi</i>	NC_002573
<i>Chara vulgaris</i>	NC_005255

## ***Results and Discussion***

### **Overall organization of *E. huxleyi* mtDNA**

The mitochondrial genome of *E. huxleyi* is a circular molecule of 29,013 bp. Figure II.1 depicts the physical and gene map of the mtDNA. The overall A+T content is 71.7%, with protein-coding regions being 72% A+T and intergenic spacers 76% A+T. This base composition is comparable to those of *Cyanidioschyzon merolae* (72.8%)(Ohta et al. 1998), *Chondrus crispus* (72.1%)(Leblanc et al. 1995), and *Reclinomonas americana* (73.9%)(Lang et al. 1997), but higher than that of *Marchantia polymorpha* (57.6%)(Oda et al. 1992) and *Arabidopsis thaliana* (55.2%)(Unseld et al. 1997). The genetic information is densely packed, with 78% of sequence specifying genes, ORFs and structural RNAs, and only 22% without detectable coding content. All the genes are encoded on the same strand suggesting that the genome is transcribed in one unit, like the mitochondrial genomes of *Monosiga brevicollis*, *Acanthamoeba castellanii*, *Dictyostelium discoideum*, *Chlamydomonas eugametos*, and *Pedinomonas minor* (Gray et al. 1998). Table II.2 lists all the genes and ORFs in the mtDNA of *E. huxleyi*. No overlapping genes were detected.

A comparison of gene order between *E. huxleyi* and *Pavlova lutheri* mtDNAs shows that no gene clusters are conserved between these two organellar genomes, although the gene content is not strikingly different (<http://megasun.bch.umontreal.ca/ogmp/projects/pluth/gen.html>). There are several features that separate the members of the class Pavlovophyceae (to which *P. lutheri* belongs) from members of the Prymnesiophyceae, including *E. huxleyi*. Chloroplast



genes (Fujiwara et al. 2001) and the 18S ribosomal DNA gene (Edwardsen et al. 2000) based phylogenies showed that members of these two groups form two distinct clades, which presumably diverged between 220 and 300 Ma (Young et al. 1992; Medlin et al. 1997). With additional complete mtDNA sequences from other members of the phylum Haptophyta and related groups of organisms, the comparative analysis of gene order pattern should be useful to understand the evolutionary changes that these genomes have undergone when the species diverged.

Intergenic regions vary in size from 1 to 2624 bp, with the majority being 1-100 bp long and only two exceeding 250 bp. The longest of these, located downstream of *trnI* and upstream of *dam*, encompasses two types of direct repeat motifs. One is a 147 bp repeat that occurs as a tandem array of 5 motifs. The other, which occurs adjacent to the first, is 246 bp long and is also arranged in a tandem array of 5 motifs. The intergenic region contains a 45 nt stem loop structure which was detected downstream of the tandem repeats and upstream of the *dam* gene (Fig II.1). The G+C content of this region is 26% and this value is approximately equal to the average of the mtDNA of *E. huxleyi* (28.3%). This region may play a role in transcription initiation or DNA replication as in the red alga *C. crispus* (Leblanc et al. 1995; Richard et al. 1998). However, no significant sequence similarity between *E. huxleyi* and *C. crispus* stem loops is discernable.

## **Gene content**

The mitochondrial genome of *E. huxleyi* codes for 21 proteins, including 14 components of the respiratory chain and 5 ribosomal proteins (Table II.2). None of



Table II.2. Genes identified in *E. huxleyi* mtDNA<sup>a</sup>

rRNA

Small subunit (1): *rns*

Large subunit (1): *rnl*

Transfer RNAs (25)

Ribosomal protein

Small subunit (4): *rps3*, *rps8*, *rps12*, *rps14*

Large subunit (1): *rpl16*

Electron transport and oxidative phosphorylation

Respiratory chain

NADH dehydrogenase (7): *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*

Ubiquinol: cytochrome c oxidoreductase (1): *cob*

Cytochrome c oxidase (3): *cox1*, *cox2*, *cox3*

ATP synthase (3): *atp4*, *atp6*, *atp9*

Other proteins

DNA adenine methyltransferase (1): *dam*

ORFs unique to *E. huxleyi* mtDNA (1)

*ORF104*

<sup>a</sup> Genes are classified according to their function. Numbers within parenthesis indicate the number of genes in a particular class. The number indexing the ORF refers to the number of amino acid residues in the deduced polypeptide.

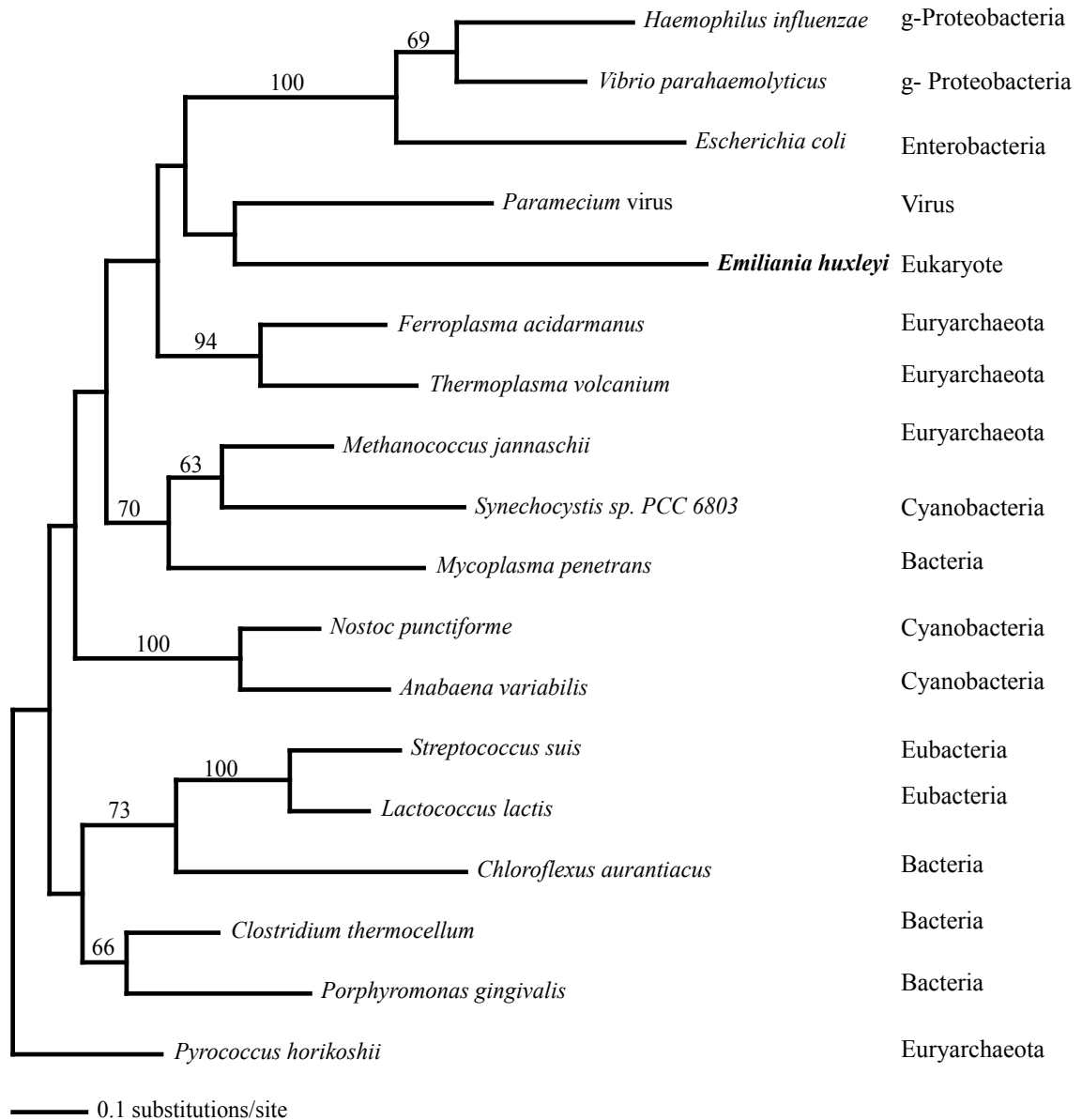
the protein coding genes contain introns, although introns are present in other mitochondrial genomes, including the haptophyte *P. lutheri*, the red alga *C. crispus* (Leblanc et al. 1995), and the liverwort *M. polymorpha* (Oda et al. 1992). Three of the genes encoding subunits of the ATP synthase complex, *atp4* (*ymf39*), *atp6*, and *atp9*, were identified. Recently, *ymf39* was identified as *atp4* (Burger et al. 2003). Seven genes encoding NADH dehydrogenase subunits, *nad1-6* and *nad4L*, were detected; these are not grouped together. In addition to these 21 protein-coding genes, a total of 27 RNA genes are present in the *E. huxleyi* mitochondrial genome, coding for 25 tRNAs and the small and large subunit (*rrs*, *rrl*) rRNAs. A 5S rRNA gene was not detected. One unique 104 aa ORF (*ORF104*) was present, and it lacks significant similarity to any entry in the public domain sequence databanks. A distinguishing organizational feature in *E. huxleyi* mtDNA is the presence of two separate coding regions that show similarity to *cox1*. The *cox1a* segment encodes the N-terminal 88 amino acid (aa) residues, and *cox1b* specifies the C-terminal 433 residues. Both segments are encoded on the same strand of the genome, interspersed with other genes, and separated in the genome by almost 10 kbp. Because only *cox1a* has a translational initiation codon and other genes are interspersed between the two *cox1* coding regions, splicing events would be required to produce the mature message. The *nad1* gene is *cis* and *trans*-spliced in plant mitochondrial genomes where the gene is split into different exons (Chapdelaine and Bonen 1991; Conklin et al. 1991; Wissinger et al. 1991). An alternative hypothesis would be that the *cox1* sequence in the mitochondrion is no longer functional, but the high degree of sequence conservation suggests that it must have been under selection until very recently.

### **A novel feature in mtDNA: adenine methyltransferase**

A unique feature of the mtDNA of *E. huxleyi* is the presence of the *dam* gene, which codes for DNA adenine methyltransferase (A-Mtase). This enzyme catalyzes the transfer of a methyl group from S-adenosyl-L-methionine (AdoMet) to the N6 position of a specific adenine in their cognate sequence (Ahmad and Rao 1996). This gene is not included in the standard set of mitochondrial genes and has not been reported for any mitochondrial genome so far studied. Significant blast hits of this gene were exclusively bacteria, viruses, and members of Archaea. A phylogenetic tree based on the *dam* gene from a number of organisms is shown in Figure II.2. A relationship of the *dam* gene from *E. huxleyi* to any other sequence is not strongly supported. The most likely tree shows *Emiliana* sister to a viral sequence (*Paramecium virus*) but with low bootstrap support. Adenine methylation plays an important role in replication, mismatch repair, and segregation of chromosomal DNA in *E. coli*, as well as regulation of gene expression and attenuation of the virulence of a number of pathogens (Heithhoff et al. 1999). The putative functional role of the *dam* gene in *E. huxleyi* could be related to mitochondrial DNA replication, modulation of gene expression or/and control of virulence of some pathogens, such as viruses. It is known that viruses infecting *E. huxleyi* can control and terminate blooms of this organism (Green and Harris 1996), and some viruses are known to target the mitochondrial genome (Hong et al. 1999).

Sequence comparisons among the members of this group of methyltransferases have revealed nine conserved motifs, of which motif I and motif

Figure II.2. Phylogenetic tree based on the *dam* gene (528 nt, excluding the third codon position). Maximum Likelihood analysis performed with PAUP\* under the GTR+I+ $\Gamma$ 4 model of evolution. Numbers above branches represent bootstrap support values (shown when > 60).



IV are highly conserved (Malone et al. 1995). These two motifs are involved in AdoMet binding and methyl group transfer. The two conserved domains of this protein are recognized in *E. huxleyi*, suggesting that it may be functional. Furthermore, the complete sequence does not suggest that it is an inactive pseudogene.

This type of methyltransferases is commonly found in Bacteria, Archaea, and viruses. N6-methylated adenine has been found in DNA of eukaryotes, such as protozoa (Rae and Spear 1978; Capowski et al. 1989), fungi (Rogers et al. 1986), higher plants (Fedoreyeva and Vanyushin 2002), and animals (Kay et al. 1994); however, the putative gene responsible for the methylation are found in the nuclear genome and correspond to a different type than the *E. huxleyi dam* gene. A nuclear encoded N6 A-Mtase isolated from wheat coleoptiles seems to be responsible for mitochondrial DNA modification that might be involved in the regulation of replication of mitochondria in plants (Fedoreyeva and Vanyushin 2002).

Correspondence analysis of the protein-coding genes in mtDNA of *E. huxleyi* (see below) revealed differences in some of the genes, including the *dam* gene. The location of the *dam* gene in the mitochondrial genome is also suggestive, since it occurs adjacent to the tandem repeats. In angiosperm mitochondria, inverted and also direct repeats appear to promote major genome rearrangements (Hanson and Folker 1992). Three alternative scenarios could account for the presence of the *dam* gene in the mtDNA of *E. huxleyi*: (1) lateral transfer of the *dam* gene from a phage or bacterial DNA; (2) vertical transmission of an ancient gene that was present in the proteobacterial progenitor of the mitochondrion; and (3) lateral transfer of the gene

from the nucleus or chloroplast to the mitochondria. To distinguish among these alternatives the distribution and phylogeny of this gene need to be examined more completely.

### **Codon usage**

Table II.3 shows the codon frequency in genes and ORFs of *E. huxleyi* mtDNA. As expected for an extremely A+T-rich genome, codons ending in A or T vastly outnumber the synonymous codons ending in G or C. The mtDNA of *E. huxleyi* does not use the standard genetic code. The codon UGA, usually serving as a translational termination signal, is used for Tryptophan (Trp). This assignment is based on protein alignments, since the codon UGA occurs where the codon for Trp is conserved in other organisms. This codon (TGA) is used preferentially, accounting for 86.7% of all Trp codons (Table II.3). This is the most common deviation from the standard translation code in mitochondria and is also found in other haptophytes (Hayashi-Ishimaru et al. 1997), *C. crispus* (Leblanc et al. 1995), *A. castellani* (Burger et al. 1995), animals, fungi, and ciliates. However, members of the order Pavlovales (phylum Haptophyta) use the universal genetic code (Inagaki et al. 1998). It is believed that the reassignment of the TGA codon to Trp occurred independently many times in the evolution of the protozoa (Inagaki et al. 1998).

No in-frame ATG start codon is present in the reading frame of *E. huxleyi* *atp4* or *rps8* genes. Protein alignments suggest that GTG may serve as the codon for translation initiation in these genes. This has been reported for the mitochondrial genome of several different organisms, such as *Porphyra purpurea* (Burger et al.

Table II.3. Codon usage in the mitochondrial genome of *E. huxleyi*<sup>a</sup>

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	482	1.62	Ser	UCU	147	1.93
	UUC	114	0.38		UCC	10	0.13
Leu	UUA	424	3.44		UCA	133	1.75
	UUG	80	0.65		UCG	42	0.55
Tyr	UAU	140	1.33	Cys	UGU	51	1.48
	UAC	71	0.67		UGC	18	0.52
ter	UAA	19	1.81	Trp	UGA	73	1.73
ter	UAG	2	0.19		UGG	11	0.26
Leu	CUU	116	0.94	Pro	CCU	73	1.73
	CUC	6	0.05		CCC	9	0.21
	CUA	99	0.80		CCA	68	1.61
	CUG	15	0.12		CCG	19	0.45
His	CAU	70	1.43	Arg	CGU	53	2.16
	CAC	28	0.57		CGC	11	0.45
Gln	CAA	109	1.85		CGA	29	1.18
	CAG	9	0.15		CGG	2	0.08
Ile	AUU	336	2.10	Thr	ACU	141	1.87
	AUC	64	0.40		ACC	11	0.15
	AUA	81	0.51		ACA	119	1.58
Met	AUG	139	1.00		ACG	31	0.41
Asn	AAU	200	1.40	Ser	AGU	95	1.25
	AAC	86	0.60		AGC	29	0.38
Lys	AAA	221	1.61	Arg	AGA	46	1.88
	AAG	53	0.39		AGG	6	0.24
Val	GUU	228	2.38	Ala	GCU	146	1.74
	GUC	17	0.18		GCC	33	0.39
	GUA	111	1.16		GCA	126	1.50
	GUG	27	0.28		GCG	31	0.37
Asp	GAU	89	1.47	Gly	GGU	152	1.92
	GAC	32	0.53		GGC	24	0.30
Glu	GAA	111	1.45		GGA	107	1.35
	GAG	42	0.55		GGG	34	0.43

<sup>a</sup> Codons are indicated in upper case letters, amino acid residues are in three-letter code; termination codons (UAA and UAG) are indicated by ter. The total number of individual codons (N) and the relative synonymous codon usage (RSCU) are shown.

1999), *Paramecium aurelia* (Pritchard et al. 1990), and *Oenothera berteriana* (Bock et al. 1994), although not referring to the same genes.

Several analyses were performed to compare the codon usage of the genes present in *E. huxleyi* mtDNA. No significant differences in codon usage, G+C content or G+C content of the silent third position were observed in identified protein coding genes or the unique ORF. However, by means of correspondence analysis of codon usage, I was able to detect that some genes were considerably different from the others, such as *atp4*, *atp9* and *dam*. The disparity of the *dam* gene was only detectable in the second axis of ordination. The G content of silent third position in *dam* gene was 20.7%, which is the highest among the protein-coding genes of *E. huxleyi* (average 11.6%). No significant differences were detected for the *ORF104* using correspondence analysis.

## **Transfer RNAs**

The tRNAs are scattered throughout the entire genome, either singly or in groups, and all lack introns. Based on the anticodon sequence of the 25 tRNAs, 24 decode the standard 20 amino acids, while one recognizes UGA, which is used for tryptophan. All the tRNA sequences can assume standard cloverleaf secondary structures, with few departures from the conventional structure, as tested with tRNAscan. This set of mitochondrial encoded tRNAs is not sufficient to decode the 62 sense codons that occur in protein-coding sequences, even when taking into account wobble and the possible modifications of their anticodons (Crick 1966). The tRNAs, which are not in the mtDNA, may be imported from the cytosol or generated



from another tRNA by partial editing or post transcriptional modification (Ohta et al. 1998). A minimum of one tRNA gene remains to be identified in order to account for the complete translation of the *E. huxleyi* mitochondrial genetic information; namely *trnL* (CAA) for leu (UUG). In addition, *trnG* (GCC) for gly (GGU and GGC) is also missing. The *trnG* (UCC) decodes GGA and GGG anticodons and is possible that it also recognizes GGU and GGC anticodons, as in *C. crispus* (Leblanc et al. 1995). Also, the *trnW* (CCA) for Trp (UGG) is not present in *E. huxleyi* mtDNA. However, *trnW* (UCA), which recognizes UGA codon as tryptophan, may also be able to decode UGG codon for the same amino acid as reported for *Tetrahymena pyriformis* mtDNA (Burger et al. 2000). The mitochondrial genome of *E. huxleyi* contains 3 tRNA genes having the methionine anticodon CAU, which include the elongator and initiator methionine-accepting mitochondrial tRNAs.

### **Phylogenetic analysis based on mitochondrial genes**

Several phylogenetic analyses have been performed based on individual and concatenated datasets. In general, individual gene analyses are congruent with each other; they find some of the major lineages monophyletic, but most branches are not strongly supported (Table II.4, Figure II.3). Concatenated analyses based on all datasets under different analytical methods recover consistently the following clades: red algae, green algae, opisthokonts, amoebzoa, heterokonts or stramenopiles, and alveolates. The placement of *C. roenbergensis* within the heterokont clade is not always strongly supported. The placement of *Emiliania huxleyi* in phylogenetic trees varies according to the method of analysis, dataset, and taxa present.

Table II.4. Bootstrap support values (when >50) found in individual gene maximum likelihood analyses under GTR+I+Γ4 using PAUP\*.

Gene Name	Number of nucleotides <sup>a</sup>	Rhodophyta monophyly	Streptophyta monophyly	Heterokonts monophyly	Opisthokonts monophyly	Relationship with haptophyte <sup>b</sup>
<i>atp6</i>	549	96	96	x	/	-
<i>atp9</i>	210	/	93	x	NA <sup>c</sup>	Green algae
<i>cob</i>	1107	100	97	88	61	Green algae
<i>cox1</i>	1455	/	/	x	/	Heterokonts
<i>cox2</i>	720	x	55	x	x	-
<i>cox3</i>	786	/	93	x	/	Heterokonts
<i>nad1</i>	717	/	/	x	x	Amoebozoa
<i>nad2</i>	396	x	x	x	x	Heterokonts
<i>nad3</i>	255	x	x	82	x	-
<i>nad4L</i>	234	x	/	x	/	-
<i>nad5</i>	1341	/	86	x	x	Green algae
<i>nad6</i>	207	/	x	x	x	Green algae
<i>rps12</i>	396	x	100	x	NA <sup>c</sup>	Red algae

<sup>a</sup> Alignment including all codon positions, but third positions were excluded from the phylogenetic analyses.

<sup>b</sup> Closest relationship to the haptophyte *Emiliana huxleyi* in the best tree.

<sup>c</sup> Opisthokonts monophyly cannot be tested because none or only one taxon is present.

Figure II.3. Individual gene phylogenetic analyses. Maximum likelihood trees under GTR+I+ $\Gamma$ 4 model of evolution using PAUP\*. Numbers above branches represent bootstrap support values.

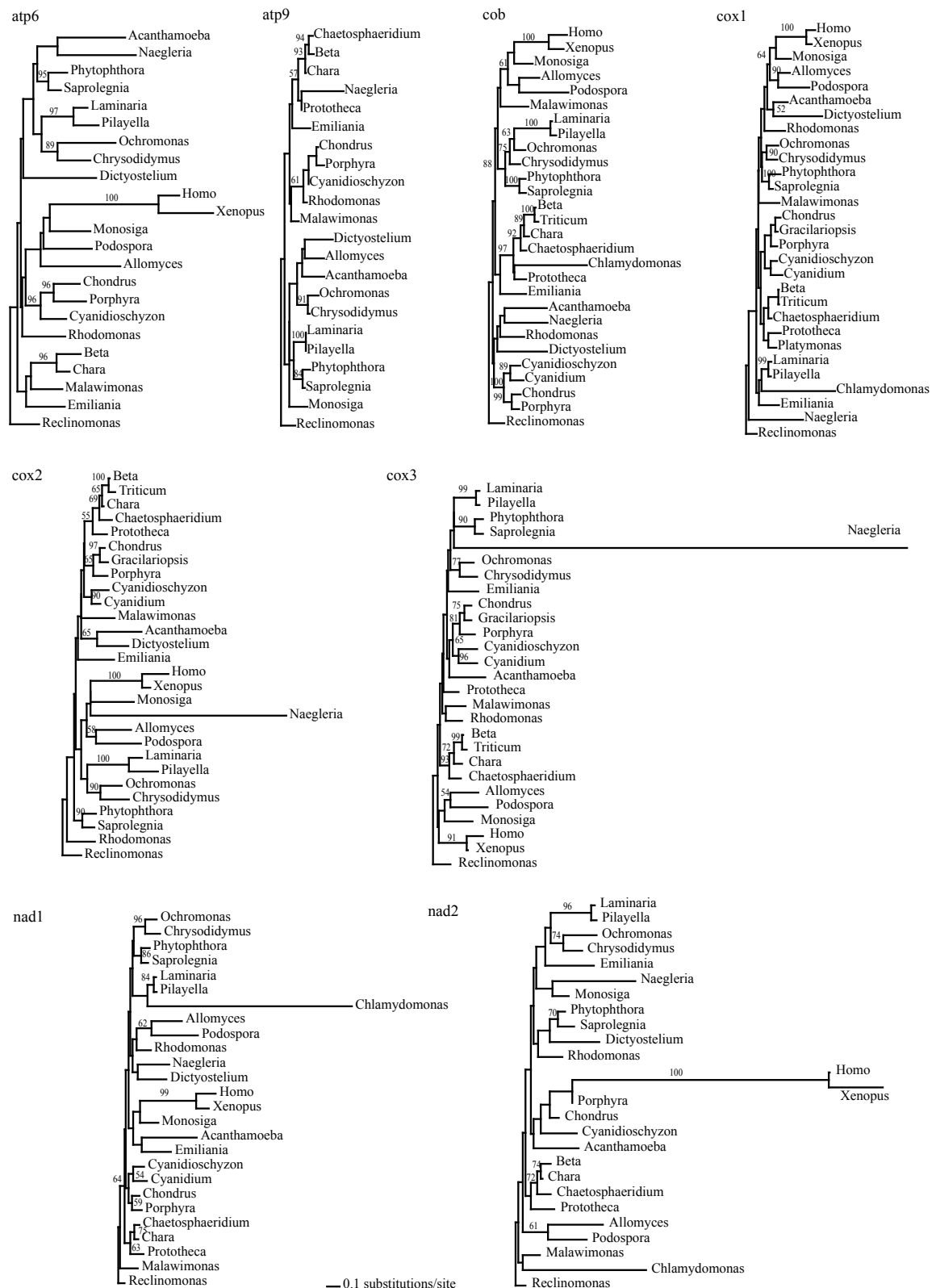
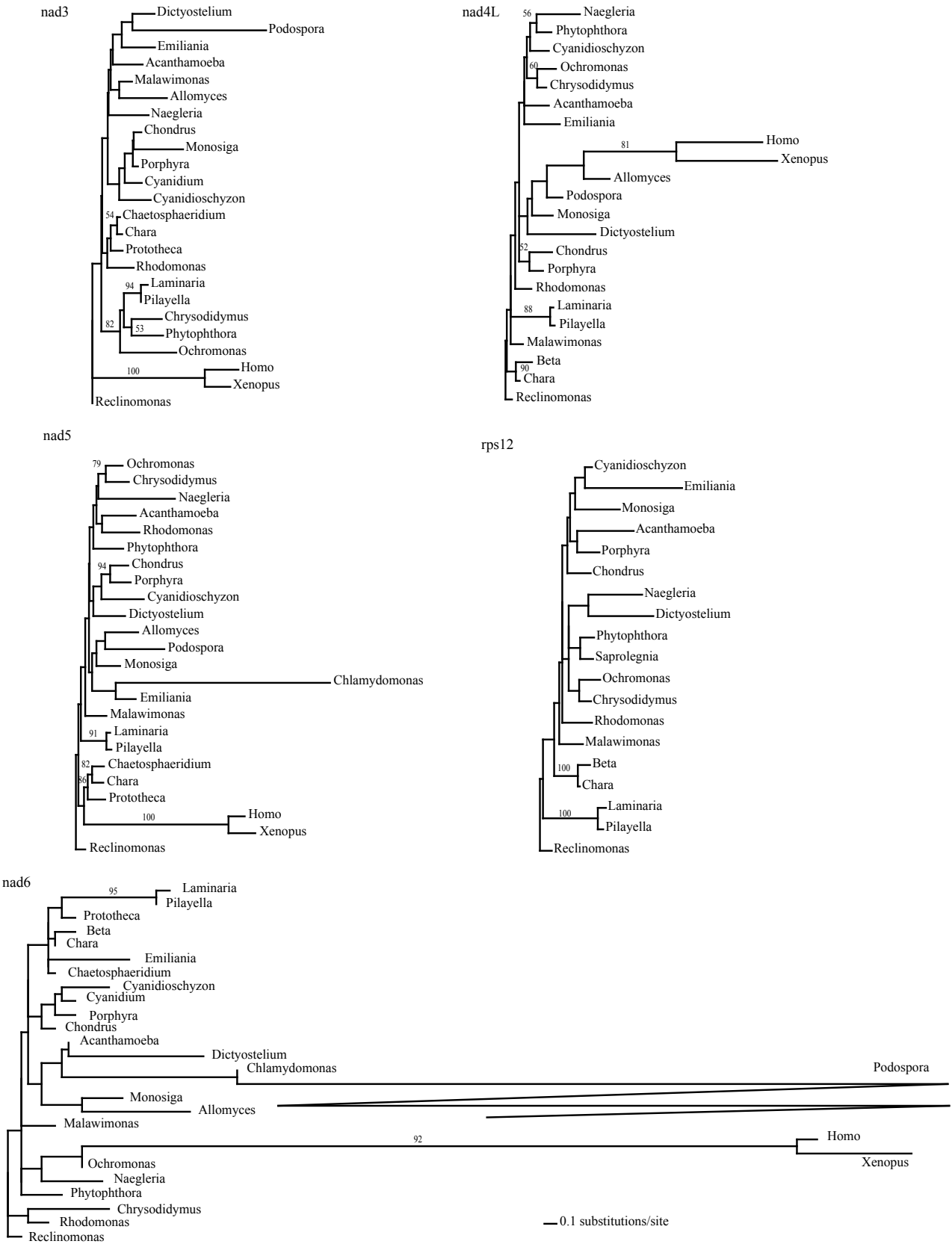


Figure II.3 (con't). Individual gene phylogenetic analyses. Maximum likelihood trees under GTR+I+Γ4 model of evolution using PAUP\*. Numbers above branches represent bootstrap support values.



Analyses based on the 4-gene dataset show *E. huxleyi* as sister to *Malawimonas jakobiformis* in ML nucleotide analysis using PAUP\* (Figure II.4), and sister to *Malawimonas* + *Rhodomonas* in ML protein analysis using PhyML (Figure II.5A). However, the Fitch-Margoliash tree using LogDet distances places *E. huxleyi* sister to the Opisthokont clade, although with weak support (Figure II.5B). When alveolates are included, *E. huxleyi* is found sister to the *C. roenbergensis* (Figure II.5C), or as a deep branching lineage sister to a diverse clade with low support (Figure II.5D). The alveolates themselves form a monophyletic group with extremely long branches that has strong support with Bayesian methods (Figure II.5D), but weak support in the ML analysis using PAUP (Figure II.5C). A Plantae clade, including Rhodophyta and Streptophyta, is recovered with variable support depending on the analytical method (Figures II.4, II.5). Overall, these trees do not identify a strong relationship of the haptophytes to heterokonts, cryptophytes, or alveolates.

Analyses based on the thirteen-gene dataset, regardless of the method used, find five monophyletic clades with variable range of support (Figure II.6). Opisthokonta, Rhodophyta, Streptophyta, and Heterokontophyta are strongly supported. Amoebozoa is monophyletic with low to strong support depending on the analysis, and it is found sister to Opisthokonts with low support. These analyses do not recover a monophyletic Plantae clade. The haptophyte *Emiliana* is found sister to the cryptophyte *Rhodomonas* but the support is low in maximum likelihood analyses based on amino acid (Figure II.6A) or nucleotide data (Figure II.6B). However, Bayesian analyses recover this grouping with high posterior probability (pp 0.98).

## **Haptophyte evolution**

Haptophytes are a monophyletic group of protists that were formerly placed with the heterokonts in the class Chrysophyceae (Pascher 1910; Bourrelly 1957), and some modern authors still emphasize a close relationship among these taxa (Andersen 1991). Ultrastructural and molecular evidence indicated that the haptophytes are a monophyletic group with the primary synapomorphy being a characteristic appendage, the haptonema (Christensen 1962; Christensen 1989; Inouye and Kawachi 1994). Studies using the nuclear genes actin and small subunit ribosomal DNA support the distinctiveness of the haptophytes, but have not shown their affinity to any other group (Bhattacharya et al. 1993; Medlin et al. 1996; Medlin et al. 1997; Tengs et al. 2000). Plastid-encoded or plastid-derived genes do show a relationship to the other chlorophyll c containing algae indicating that the plastids of haptophytes, heterokonts, dinoflagellates, and cryptophytes form a monophyletic group (Yoon et al. 2002b; Harper and Keeling 2003; Bachvaroff et al. 2005). However, these data are not informative for relationships among cytosolic genomes.

Although phylogenetic analyses clearly support a red algal origin of chlorophyll c containing (chl c) plastids, the number of events that gave rise to them and the phylogenetic relationships among the host cells remains unclear. There are two competing hypotheses regarding the number of endosymbiotic events in the chl c algae. In the first of these, a single endosymbiotic event for all the chl c algae would imply that the host cells are also monophyletic (Cavalier-Smith 1981; Cavalier-Smith 1989; Yoon et al. 2002b), and that plastid-loss has occurred in the non-photosynthetic lineages. The clade postulated by this hypothesis is known as chromalveolates

Figure II.4. Maximum likelihood tree inferred from the first two codon positions of a concatenated *cob*, *cox1*, *cox2*, and *cox3* alignment using PAUP\* with a GTR + I +  $\Gamma$ 4 model of sequence evolution. The numbers above the branches indicate the bootstrap proportion using the first two codon positions, the bootstrap proportion when all positions are used, and the Bayesian posterior probability of the branch when all positions are used with the GTR + I +  $\Gamma$ 4 model, where the same branches were recovered. Thicker branches indicate 100% bootstrap support with both datasets, and a posterior probability of 1.0.

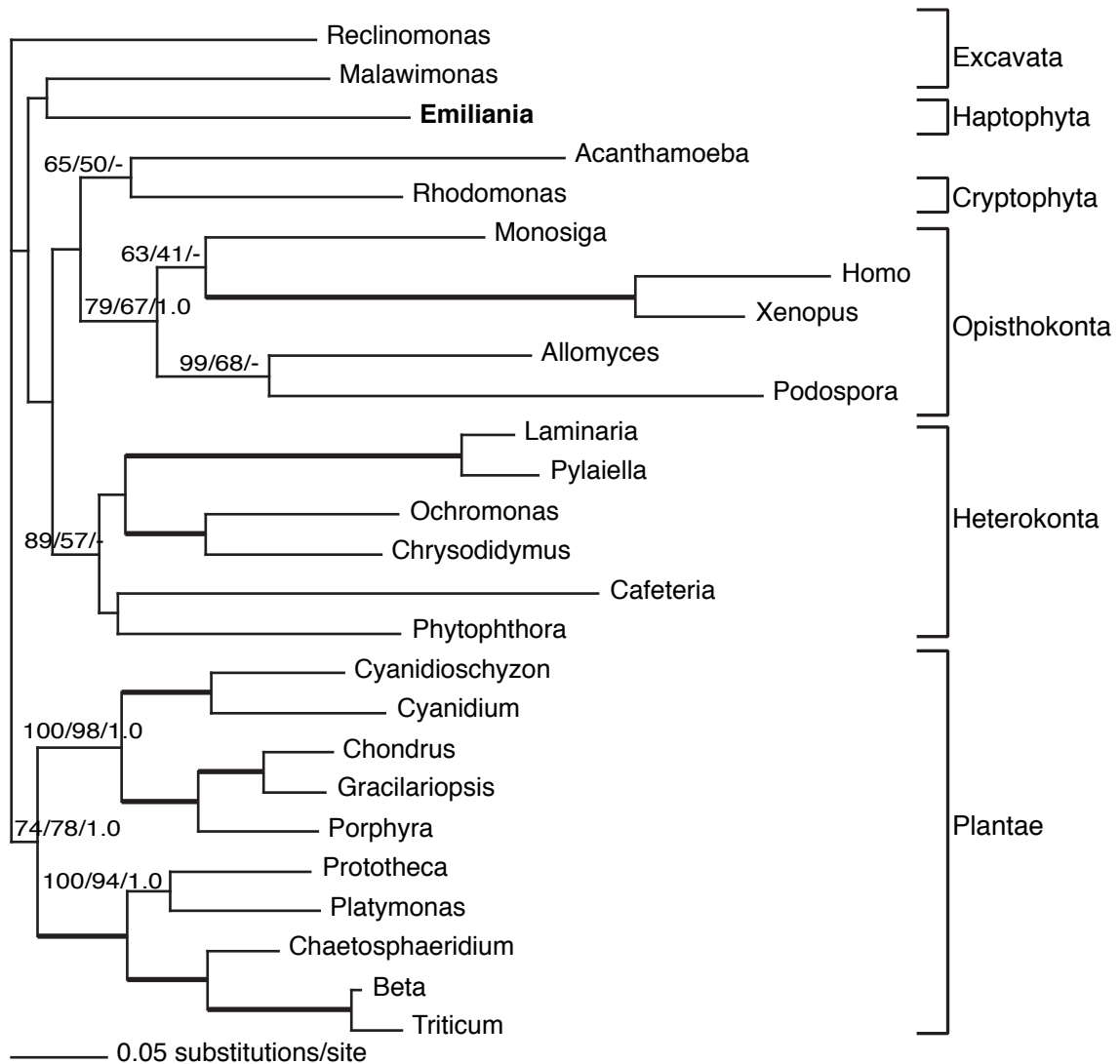


Figure II.5. Phylogenetic analyses based on the genes *cob*, *cox1*, *cox2*, and *cox3*. A. Maximum Likelihood tree based on protein data using PhyML under JTT model of evolution. Numbers represent bootstrap values. B. MrBayes tree based on GTR+I+ $\Gamma$ 4 model, excluding the third codon position, with posterior probabilities above the branches. Numbers below branches are bootstrap values from a LogDet distance analysis, when > 60. C and D. Phylogenetic analyses based on nucleotides under GTR+I+ $\Gamma$ 4, excluding the third codon position and including Alveolates. C. Maximum Likelihood tree, showing bootstrap values above branches. D. MrBayes tree with posterior probabilities above branches.

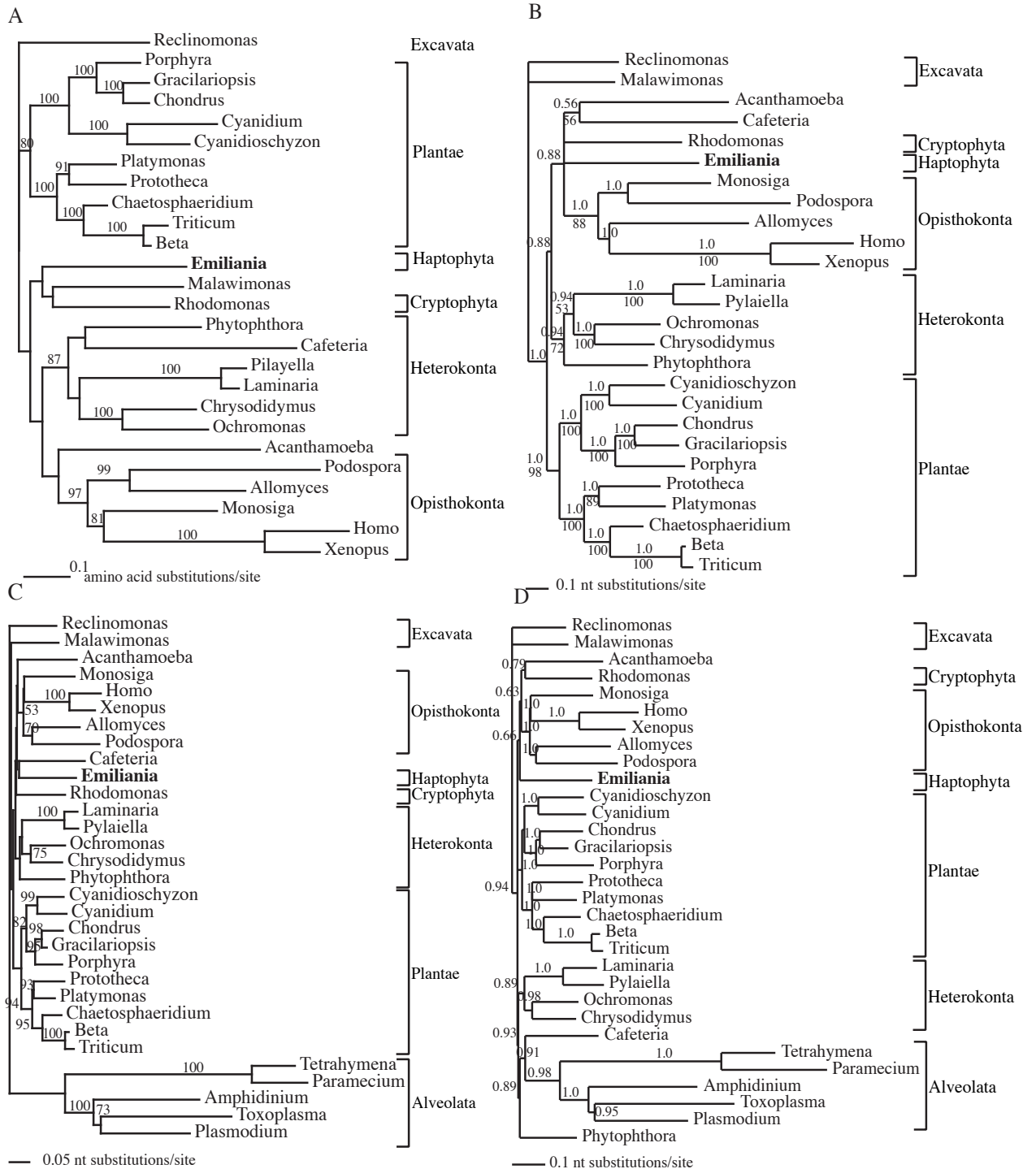
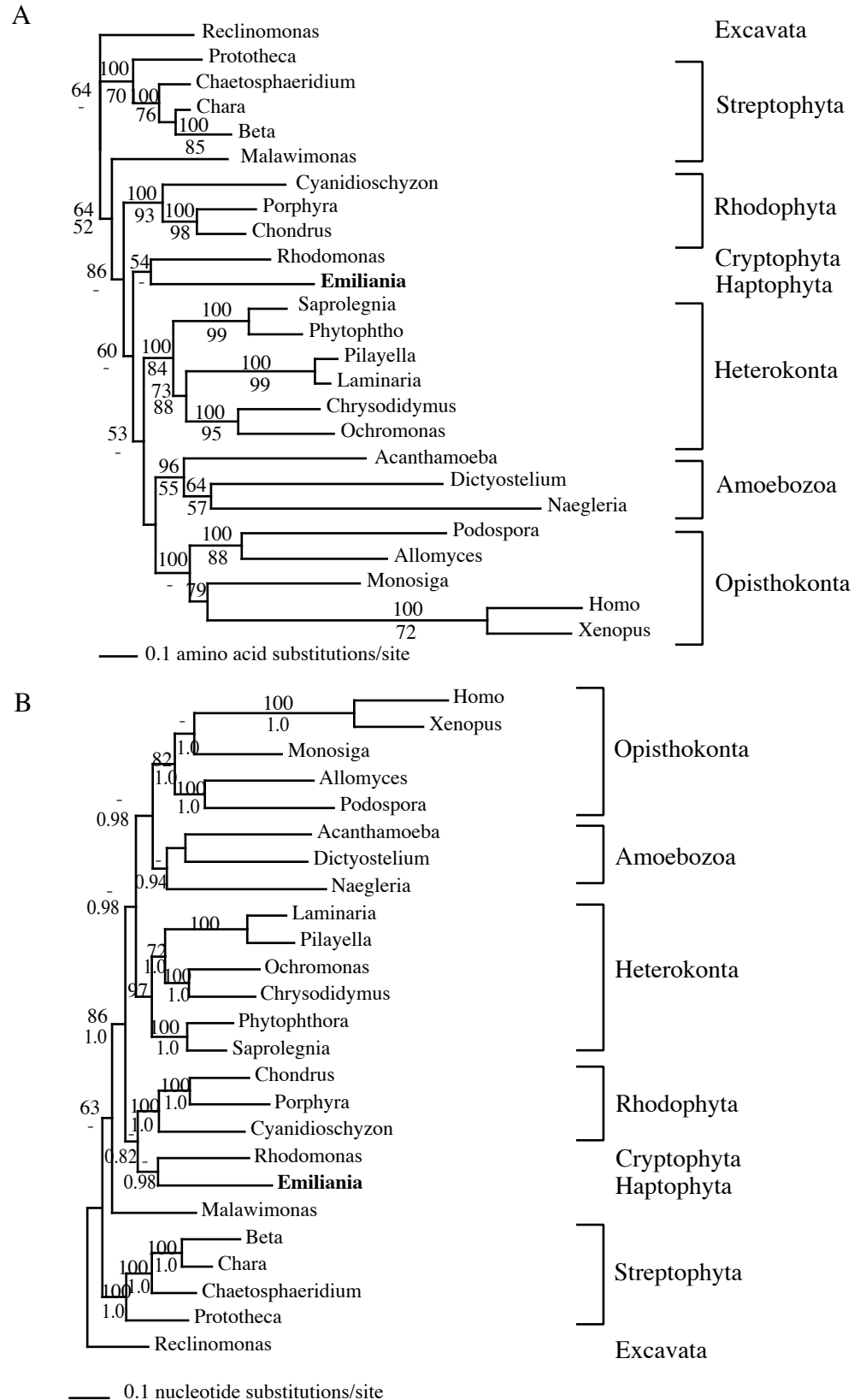




Figure II.6. Phylogenetic analyses based on 13 mitochondrial genes. Support values shown when > 50. A. Maximum likelihood (ML) tree based on proteins under JTT model of evolution with PhyML. Numbers above branches are bootstrap values from the ML analysis; numbers below branches are quartet puzzling support values from TreePuzzle analysis. B. ML tree based on nucleotides (excluding third position) using PAUP\* under GTR+I+ $\Gamma$ 4 model. Numbers above branches are bootstrap values. Number below branches correspond to posterior probabilities from Bayesian analysis.



(Cavalier-Smith 1999; Cavalier-Smith 2002). An alternative hypothesis would infer multiple endosymbiotic events in separate chl c lineages, would allow host and endosymbiont phylogenies to be incongruent (Cavalier-Smith et al. 1994), and would imply a non-photosynthetic ancestor for each lineage. Several intermediate hypotheses could also be proposed. However, most authors seem to agree that the question remains unresolved. The analyses presented here examined the host cell phylogenetic relationships using concatenated mitochondrial genes in order to assess the monophyly of the chromalveolate clade. Mitochondrial data can be used to resolve relationships among eukaryotes, such as the monophyletic origin of red and green primary plastids and can complement data from the nuclear genome. Like previous analyses of mitochondrial data (Burger et al. 1999), the 4-gene dataset supports a single origin of the primary plastid in the Plantae. The hypothesis of a single origin of the chl c containing hosts is not supported by these data, but the bootstrap and posterior probabilities are weak, leaving open the possibility that this hypothesis is correct. *E. huxleyi* is clearly excluded from the heterokonts in these analyses, but its placement as sister taxon to the heterokonts cannot be rejected. Concatenated mitochondrial data presented here do not strongly resolve the relationships among chl c containing algae, indicating that further study of relationships among these taxa is needed. The identity of the sibling taxon to haptophytes remains an unsolved problem, although these analyses suggest a relationship with cryptophytes.

# **Chapter III – The Complete Plastid Genome Sequence of the Haptophyte *Emiliana huxleyi*: a Comparison to Other Plastid Genomes**

## ***Abstract***

The complete nucleotide sequence of the plastid genome of the haptophyte *Emiliana huxleyi* has been determined. *E. huxleyi* is the most abundant coccolithophorid and has a key role in the carbon cycle. It is also implicated in the production of dimethylsulphide (DMS), which is involved in cloud nucleation and may affect the global climate. Here, I report the plastid genome sequence of this ecologically and economically important species and compare its content and arrangement to other known plastid genomes. The genome is circular and consists of 105,309 bp with two inverted repeats of 4,841 bp each. In terms of both genome size and gene content *E. huxleyi* cpDNA is substantially smaller than any other from the red plastid lineage. The genetic information is densely packed, with 86.8% of the genome specifying 110 identified protein-coding genes, 9 open reading frames, 30 tRNAs, and 6 rRNAs. No introns were present and only two overlapping genes (*psbD*, *psbC*) were found. The GC content of the genome is 36.8%. Parsimony analysis based on presence/absence of 261 plastid-encoded genes in 18 plastid genomes was performed. A detailed comparison to other plastid genomes, based on gene content, gene function, and gene cluster analysis is discussed. These analyses suggest a close relationship of the *E. huxleyi* cpDNA to the chlorophyll c containing

plastids from heterokonts and cryptophytes, and they support the origin of the chlorophyll c containing plastids from the red algal lineage.

## ***Introduction***

Oxygenic photosynthesis arose once in the cyanobacteria. Later in evolution, several unrelated lineages of non-photosynthetic eukaryotes acquired this capability by engulfing a photosynthetic organism and keeping it as a permanent endosymbiont. In modern organisms, the highly reduced descendants of these endosymbionts are the cellular organelles known as plastids (or, in green plants, chloroplasts). Depending upon the group in question, plastids were acquired through a process of primary, secondary, or tertiary endosymbiosis (Delwiche and Palmer 1997). Primary endosymbiosis is the process by which a photosynthetic prokaryote (cyanobacterium) was engulfed and integrated into a non-photosynthetic eukaryotic host cell. Three photosynthetic eukaryotic lineages, namely red algae, green algae, and glaucophytes, possess primary plastids, all of which were probably derived from a single endosymbiotic event (Van der Auwera et al. 1998; Cavalier-Smith 2000; Moreira et al. 2000). In secondary endosymbiosis, non-photosynthetic eukaryotic host cells acquired a photosynthetic eukaryote from the green or red algal lineages. Four photosynthetic lineages, including haptophytes, dinoflagellates, cryptophytes and heterokonts, contain secondary plastids with chlorophyll c (chl c) as a main photosynthetic pigment (Gibbs 1981a; Palmer and Delwiche 1996; Delwiche and Palmer 1997). These plastids are thought to have been acquired from the red algal lineage (Delwiche and Palmer 1997; Medlin et al. 1997; Durnford et al. 1999; Yoon

Table III.1. List of plastid genomes from different photosynthetic eukaryotes published as of March 2005.

Taxon	Accession number	Taxonomy	Type of plastid	Length (bp)	Number of proteins
<i>Chlamydomonas reinhardtii</i>	NC_005353	Chlorophyta	1°, green	203,828	69
<i>Chlorella vulgaris</i>	NC_001865	Chlorophyta	1°, green	150,613	175
<i>Nephroselmis olivacea</i>	NC_000927	Chlorophyta	1°, green	200,799	155
<i>Adiantum capillus-veneris</i>	NC_004766	Streptophyta	1°, green	150,568	88
<i>Amborella trichopoda</i>	NC_005086	Streptophyta	1°, green	162,668	86
<i>Anthoceros formosae</i>	NC_004543	Streptophyta	1°, green	161,162	92
<i>Arabidopsis thaliana</i>	NC_000932	Streptophyta	1°, green	154,478	88
<i>Atropa belladonna</i>	NC_004561	Streptophyta	1°, green	156,687	87
<i>Calycanthus floridus</i>	NC_004993	Streptophyta	1°, green	153,337	88
<i>Chaetosphaeridium globosum</i>	NC_004115	Streptophyta	1°, green	131,183	98
<i>Lotus corniculatus</i>	NC_002694	Streptophyta	1°, green	150,519	83
<i>Marchantia polymorpha</i>	NC_001319	Streptophyta	1°, green	121,024	89
<i>Mesostigma viride</i>	NC_002186	Streptophyta	1°, green	118,360	105
<i>Nicotiana tabacum</i>	NC_001879	Streptophyta	1°, green	155,939	102
<i>Nymphaea alba</i>	NC_006050	Streptophyta	1°, green	159,930	89
<i>Oenothera elata</i>	NC_002694	Streptophyta	1°, green	163,935	118
<i>Oryza nivara</i>	NC_005973	Streptophyta	1°, green	134,494	119
<i>Oryza sativa</i>	NC_001320	Streptophyta	1°, green	134,525	108
<i>Panax schinseng</i>	NC_006290	Streptophyta	1°, green	156,318	80
<i>Physcomitrella patens</i>	NC_005087	Streptophyta	1°, green	122,890	85
<i>Pinus koraiensis</i>	NC_004677	Streptophyta	1°, green	116,866	164
<i>Pinus thunbergii</i>	NC_001631	Streptophyta	1°, green	119,707	159
<i>Psilotum nudum</i>	NC_003386	Streptophyta	1°, green	138,829	101
<i>Saccharum officinarum</i>	NC_006084	Streptophyta	1°, green	141,182	117
<i>Spinacia oleracea</i>	NC_002202	Streptophyta	1°, green	150,725	100
<i>Triticum aestivum</i>	NC_002762	Streptophyta	1°, green	134,545	84
<i>Zea mays</i>	NC_001666	Streptophyta	1°, green	140,387	111
<i>Euglena gracilis</i>	NC_001603	Euglenophyta	2°, green	143,172	66
<i>Cyanophora paradoxa</i>	NC_001675	Glaucophyta	1°, glaucophyte	135,599	149
<i>Cyanidioschyzon merolae</i>	NC_004799	Rhodophyta	1°, red	149,987	207
<i>Cyanidium caldarium</i>	NC_001840	Rhodophyta	1°, red	164,921	198
<i>Gracilaria tenuistipitata</i>	NC_006137	Rhodophyta	1°, red	183,883	203
<i>Porphyra purpurea</i>	NC_000925	Rhodophyta	1°, red	191,028	209
<i>Guillardia theta</i>	NC_000926	Cryptophyta	2°, red	121,524	147
<i>Odontella sinensis</i>	NC_001713	Heterokontophyta	2°, red?	119,704	140
<i>Emiliania huxleyi</i>	AY741371	Haptophyta	2°, red?	105,309	119

et al. 2002b), and I will refer to them together with the plastids from the red algae as “red plastids”.

During the process of organelle genome reduction, most of the genes of the endosymbiont were transferred to the nuclear genome of the host, while many others were lost and only a few remained as the plastid genome (Bachvaroff et al. 2004; Hackett et al. 2004). More than thirty “green plastid” genomes of members of Chlorophyta and Embryophyta, including secondary plastids of Euglenoids, have been sequenced. Only six plastid genomes from the red plastids (four from red algae, and two from chl c containing algae) have been sequenced to date (Table III.1). The plastid genome of dinoflagellates has been characterized in different peridinin-containing dinoflagellates, although they are far from being understood. Plastid genes in dinoflagellates have been found to be located in the plastid (Takishita et al. 2003) or in the nucleus (Laatsch et al. 2004) in minicircles containing only 1-3 genes. The last major group of algae whose plastid genome has remained essentially uncharacterized until now are the haptophytes.

The haptophyte *Emiliania huxleyi* is the most abundant of the coccolithophorids, and one of the dominant marine calcifying phytoplankton species. Coccolithophorids are responsible for a significant amount of biogenic carbonate precipitation and are among the major contributors to marine primary production (Falkowski et al. 2000; Riebesell et al. 2000). Despite its ecological, economical, and evolutionary importance (Brown and Yoder 1994; Walsh and Mann 1995; Malin and Kirst 1997), until recently little was known about the molecular genetics of this alga. Lately, however, *E. huxleyi* has been the focus of several genomic studies, including

an EST project that emphasized the molecular mechanisms of biomineralization and coccolithogenesis (Wahlund et al. 2004), the complete sequence of the mitochondrial genome including analyses of the phylogenetic relationships of this alga (Sanchez-Puerta et al. 2004), and an ongoing nuclear genome sequencing project of this species (Betsy Read pers. comm.). This chapter (published as Sanchez-Puerta et al. 2005) presents the complete sequence of the chloroplast DNA (cpDNA) of the haptophyte *Emiliania huxleyi* (Lohmann) Hay & Mohler, description of the main features and comparison of gene content, order, and functions to other plastid genomes. The complete nucleotide sequence and study of the plastid genome of the haptophyte *E. huxleyi* is important to have a better understanding of plastid genome evolution, to learn more about the pattern of gene transfer to the nucleus in photosynthetic organisms, to examine the genetic properties of the plastid genome, and to investigate the evolution of the chl c containing plastid and host cells.

### ***Materials and Methods***

The complete cpDNA sequence of *Emiliania huxleyi* has been deposited in GenBank (AY741371).

### **Culture of *E. huxleyi* and cpDNA isolation**

The axenic strain of *E. huxleyi* was obtained from Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP # 373). Cells were grown in Guillard's f/2 medium (Andersen et al. 1997) at 17°C with a 14h/10h L:D cycle. The

Table III.2: List of genes used for the maximum parsimony analysis

<i>accA</i>	<i>dsbD</i>	<i>petA</i>	<i>rne</i>	<i>rps9</i>	<i>ycf29</i>	<i>ycf73</i>
<i>accB</i>	<i>fabH</i>	<i>petB</i>	<i>rpl1</i>	<i>rps10</i>	<i>ycf32</i>	<i>ycf74</i>
<i>accD</i>	<i>fdx</i>	<i>petD</i>	<i>rpl2</i>	<i>rps11</i>	<i>ycf33</i>	<i>ycf75</i>
<i>acpP/A</i>	<i>ftrB</i>	<i>petF</i>	<i>rpl3</i>	<i>rps12</i>	<i>ycf34</i>	<i>ycf76</i>
<i>apcA</i>	<i>ftsH</i>	<i>petG</i>	<i>rpl4</i>	<i>rps13</i>	<i>ycf35</i>	<i>ycf77</i>
<i>apcB</i>	<i>ftsW</i>	<i>petJ</i>	<i>rpl5</i>	<i>rps14</i>	<i>ycf36</i>	<i>ycf78</i>
<i>apcD</i>	<i>glnB</i>	<i>petL</i>	<i>rpl6</i>	<i>rps15</i>	<i>ycf37</i>	<i>ycf80</i>
<i>apcE</i>	<i>gltB</i>	<i>petM</i>	<i>rpl9</i>	<i>rps16</i>	<i>ycf38</i>	<i>ycf81</i>
<i>apcF</i>	<i>groEL</i>	<i>pgmA</i>	<i>rpl11</i>	<i>rps17</i>	<i>ycf39</i>	<i>ycf82</i>
<i>argB</i>	<i>groES</i>	<i>preA</i>	<i>rpl12</i>	<i>rps18</i>	<i>ycf40</i>	<i>ycf83</i>
<i>atpA</i>	<i>hemA</i>	<i>psaA</i>	<i>rpl13</i>	<i>rps19</i>	<i>ycf41</i>	<i>ycf84</i>
<i>atpB</i>	<i>hisH</i>	<i>psaB</i>	<i>rpl14</i>	<i>rps20</i>	<i>ycf43</i>	<i>ycf85</i>
<i>atpD</i>	<i>I-CvuI</i>	<i>psaC</i>	<i>rpl16</i>	<i>secA</i>	<i>ycf44</i>	<i>ycf86</i>
<i>atpE</i>	<i>ilvB</i>	<i>psaD</i>	<i>rpl18</i>	<i>secY</i>	<i>ycf45</i>	
<i>atpF</i>	<i>ilvH</i>	<i>psaE</i>	<i>rpl19</i>	<i>sufB</i>	<i>ycf46</i>	
<i>atpG</i>	<i>infA</i>	<i>psaF</i>	<i>rpl20</i>	<i>syfB</i>	<i>ycf47</i>	
<i>atpH</i>	<i>infB</i>	<i>psaI</i>	<i>rpl21</i>	<i>syh</i>	<i>ycf48</i>	
<i>atpI</i>	<i>infC</i>	<i>psaJ</i>	<i>rpl22</i>	<i>thiG</i>	<i>ycf49</i>	
<i>basI</i>	<i>matK</i>	<i>psaK</i>	<i>rpl23</i>	<i>trpA</i>	<i>ycf50</i>	
<i>bioY</i>	<i>minD</i>	<i>psaL</i>	<i>rpl24</i>	<i>trpG</i>	<i>ycf51</i>	
<i>carA</i>	<i>minE</i>	<i>psaM</i>	<i>rpl27</i>	<i>trxA</i>	<i>ycf52</i>	
<i>cfxQ</i>	<i>mntA</i>	<i>psbA</i>	<i>rpl28</i>	<i>tsf</i>	<i>ycf53</i>	
<i>ccsA</i>	<i>mntB</i>	<i>psbB</i>	<i>rpl29</i>	<i>tufA</i>	<i>ycf54</i>	
<i>cemA</i>	<i>moeB</i>	<i>psbC</i>	<i>rpl31</i>	<i>upp</i>	<i>ycf55</i>	
<i>chlB</i>	<i>nadA</i>	<i>psbD</i>	<i>rpl32</i>	<i>psbW</i>	<i>ycf56</i>	
<i>chlI</i>	<i>nblA</i>	<i>psbE</i>	<i>rpl33</i>	<i>ycf1</i>	<i>ycf57</i>	
<i>chlL</i>	<i>ndhA</i>	<i>psbF</i>	<i>rpl34</i>	<i>ycf2</i>	<i>ycf58</i>	
<i>chlN</i>	<i>ndhB</i>	<i>psbH</i>	<i>rpl35</i>	<i>ycf3</i>	<i>ycf59</i>	
<i>clpC</i>	<i>ndhC</i>	<i>psbI</i>	<i>rpl36</i>	<i>ycf4</i>	<i>ycf60</i>	
<i>clpP</i>	<i>ndhD</i>	<i>psbJ</i>	<i>rpoA</i>	<i>ycf6</i>	<i>ycf61</i>	
<i>cpcA</i>	<i>ndhE</i>	<i>psbK</i>	<i>rpoB</i>	<i>ycf12</i>	<i>ycf62</i>	
<i>cpcB</i>	<i>ndhF</i>	<i>psbL</i>	<i>rpoC1</i>	<i>ycf13</i>	<i>ycf63</i>	
<i>cpcG</i>	<i>ndhG</i>	<i>psbM</i>	<i>rpoC2</i>	<i>ycf15</i>	<i>ycf64</i>	
<i>cpeA</i>	<i>ndhH</i>	<i>psbN</i>	<i>rps1</i>	<i>ycf16</i>	<i>ycf65</i>	
<i>cpeB</i>	<i>ndhI</i>	<i>psbT</i>	<i>rps2</i>	<i>ycf17</i>	<i>ycf66</i>	
<i>crtE</i>	<i>ndhJ</i>	<i>psbV</i>	<i>rps3</i>	<i>ycf19</i>	<i>ycf67</i>	
<i>cysA</i>	<i>ndhK</i>	<i>psbX</i>	<i>rps4</i>	<i>ycf20</i>	<i>ycf68</i>	
<i>cysT</i>	<i>ntcA</i>	<i>rbcL(g,r)</i>	<i>rps5</i>	<i>ycf21</i>	<i>ycf69</i>	
<i>dfr</i>	<i>odpA</i>	<i>rbcR</i>	<i>rps6</i>	<i>ycf22</i>	<i>ycf70</i>	
<i>dnaB</i>	<i>odpB</i>	<i>rbcS(g,r)</i>	<i>rps7</i>	<i>ycf23</i>	<i>ycf71</i>	
<i>dnaK</i>	<i>psbA</i>	<i>rdpO</i>	<i>rps8</i>	<i>ycf27</i>	<i>ycf72</i>	



plastid DNA was isolated according to previously described methods (Chesnick and Cattolico 1993; Sanchez-Puerta et al. 2004).

### **Library construction and DNA sequencing**

Two overlapping libraries were constructed by digesting plastid DNA with the restriction endonuclease *HindIII* and *EcoRI*. The resulting fragments were cloned in pGEM -3Zf(+) (Promega, WI) using *Escherichia coli* XL-10 Gold Ultracompetent Cells (Stratagene, CA) as the host bacterium. Plasmids from individual clones were isolated using the ‘miniprep’ procedure (Sambrook and Russell 2001), and sequenced using dye terminator chemistry (ABI). The M13-20 primer was used for 5’ and T7 primer for 3’ sequencing. Primer walking was used to determine the full sequence of longer clones and to obtain double stranded sequencing reads. The polymerase chain reaction was used to order the fragments, fill gaps and obtain double stranded coverage.

### **Data analysis**

Sequences were edited using the program Sequencher (GeneCodes Corp., MI). Vector and low quality bases were removed, and manual editing was performed. Sequence reads were assembled using the contig assembly function of Sequencher.

The annotation of the genome was performed in part using DOGMA (Wyman et al. 2004). Putative open reading frames (ORFs) were identified by performing BLAST searches of the GenBank databases at the National Center for Biotechnology Information (NCBI). Transfer RNAs were detected with tRNAscan SE Search Server (Lowe and Eddy 1997). Analysis of codon usage by genes was calculated using

CodonW (University of Nottingham; <http://bioweb.pasteur.fr/seanal/interfaces/codonw.html>). Annotated plastid genomes of other organisms were obtained from NCBI Entrez-Genome database (Table III.1), and used for comparisons.

### **Cladistic analysis**

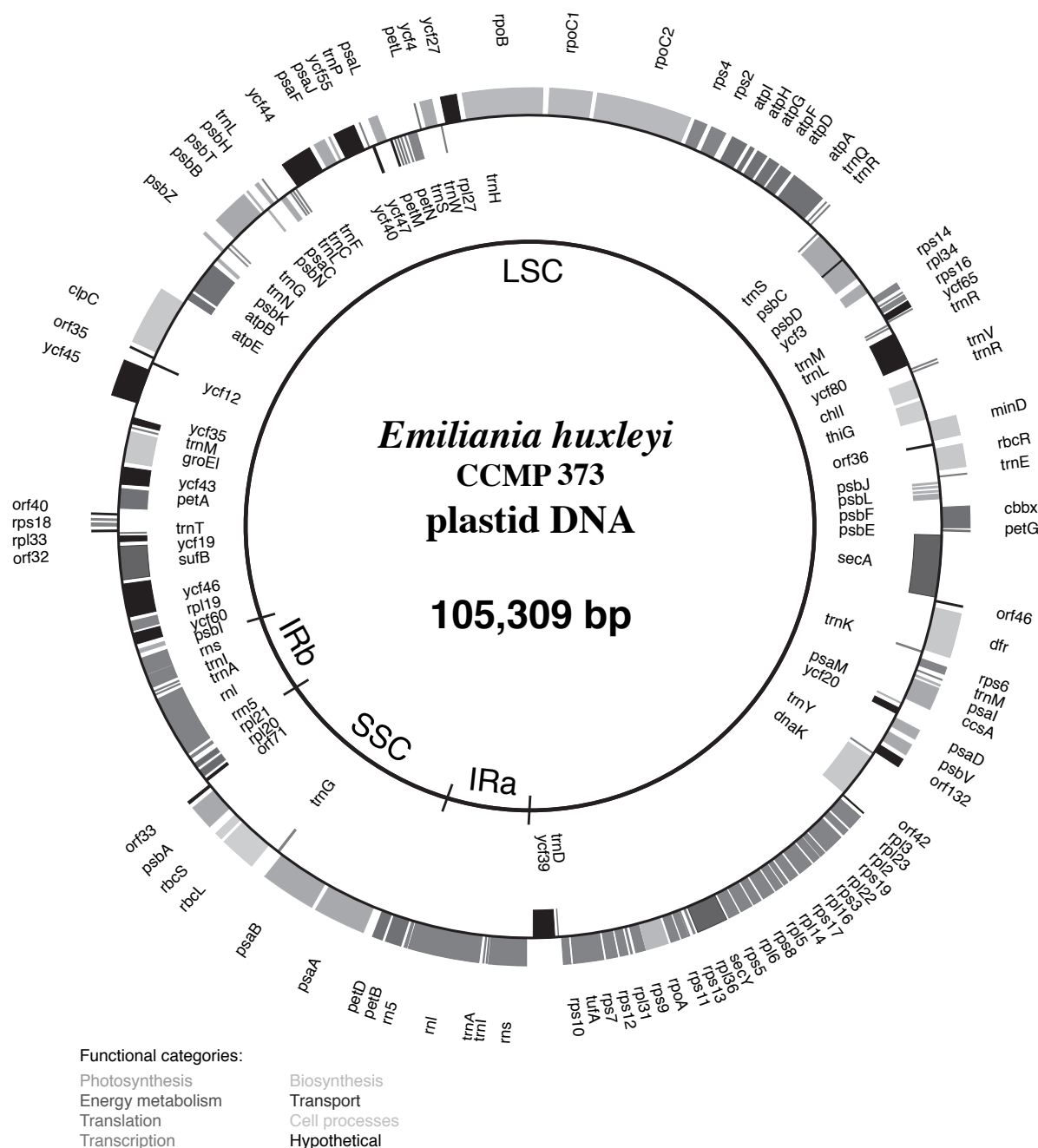
I performed parsimony analysis based on presence and absence of 261 genes (Table III.2) in 18 plastid genomes from diverse hosts, and the genome of the cyanobacterium *Synechocystis* sp. as the outgroup. The tree search was carried out using branch and bound algorithm in PAUP\*4b10 (Swofford et al. 2002). The loss of plastid genes from the plastid genome was considered a derived character, and the step matrix followed an irreversible model, under the Camin-Sokal parsimony criterion. Bootstrap support values were calculated from 1,000 heuristic searches with random stepwise addition.

## ***Results and Discussion***

### **Overall organization of *E. huxleyi* cpDNA**

The plastid genome of *E. huxleyi* consists of a circular molecule of 105,309 bp and it represents the smallest plastid genome among the red plastids so far studied. Figure III.1 depicts the physical and gene map of the cpDNA. The plastid genome contains two inverted repeats of 4,841 bp each, including genes encoding for three subunits of ribosomal RNA (16S, 23S, and 5S rRNA) and two transfer RNAs (*trnI*, *trnA*). A small single copy (SSC) and a large single copy (LSC) of 11,183 bp and

Figure III.1. Physical map and gene organization of the *E. huxleyi* plastid genome. Genes drawn outside the circle are transcribed clockwise. Gene abbreviations are listed in Table III.4. Small single copy (SCC), large single copy (LSC), and inverted repeats (IRa, IRb) are indicated. The origin for numbering of the GenBank record (O) is the first nucleotide of IRa, and numbering proceeds clockwise. Genes are colored depending on their functional categories as indicated.



84,444 bp, respectively, separate the two repeats. The 23S rDNA gene is closest to the SSC, as is the case in the plastid genomes of most land plants (Kolodner and Twarei 1979; Maier et al. 1995; Goremykin et al. 2004; Kim and Lee 2004), the heterokont *Odontella sinensis* (Kowallik et al. 1995), and the cryptophyte *Guillardia theta* (Douglas and Penny 1999), but opposite to the plastid genome arrangement of the fern *Adiantum capillus-veneris* (Wolf et al. 2003). The inverted repeats differ in three nucleotides located in non-coding regions. Inverted repeats are present in the cpDNAs of some members of the red algae, such as *Porphyra yezoensis* (Shivji 1991), but are absent in many others, such as *Porphyra purpurea* (Reith and Munholland 1995), *Cyanidioschyzon merolae* (Ohta et al. 2003), and *Cyanidium caldarium* (Glöckner et al. 2000) cpDNAs, which contain direct repeats. All known chromophyte plastid genomes contain inverted repeats. *O. sinensis* contains two inverted repeats of 7,725 bp, which include the rRNA operon and the genes *ycf32* and *rpl32* (Kowallik et al. 1995), while *G. theta* contains two inverted repeats of ca. 4,900 bp including only the rRNA operon (Douglas and Penny 1999). These data raise the question of whether the presence of an inverted repeat is ancestral in the red algal lineage and in the chromophyte plastids. To answer this question will require more cpDNA data from red algae and chromophytes. Furthermore, accurate reconstruction of such ancestral character states will only be reliable when the red algal lineage(s) that were the source of the chromophyte plastids have been identified with confidence.

The overall G+C content is 36.8%, with protein-coding regions being 37.4 % G+C and intergenic spacers 24% G+C. This base composition is comparable to that

of *C. merolae* (37.6%)(Ohta et al. 2003), and *G. theta* (33%)(Douglas and Penny 1999), but higher than that of *Chaetosphaeridium globosum* (29.6%)(Turmel et al. 2002) and *Physcomitrella patens* (28.5%)(Sugiura et al. 2003). The genetic information is densely packed, with 86.8% of sequence specifying genes, ORFs and structural RNAs, and only 13.2% without detectable coding content. Intergenic regions vary in size from 1 to 349 bp, with the majority being 1-100 bp long. The distribution of intergenic regions is similar to that of the plastid genomes of *G. theta* (Douglas and Penny 1999) and *O. sinensis* (Kowallik et al. 1995). Only one case of overlapping genes was detected, in contrast to the high number of overlapping genes in the plastid genome of the red alga *C. merolae* (Ohta et al. 2003). The genes *psbC* and *psbD* in *E. huxleyi* cpDNA share an overlapping region of 52 bp. The same two genes overlap in the plastid genomes of the red alga *Gracilaria tenuistipitata* (by 92 bp), in *G. theta* (by 94 bp), and *O. sinensis* (by 52 bp).

### **Codon usage and transfer RNA genes**

The cpDNA of *E. huxleyi* uses the standard genetic code, with three translation termination codons (TAA, TAG, TGA). Table III.3 shows the codon frequency in genes and ORFs of *E. huxleyi* cpDNA. As expected for an extremely A+T-rich genome, codons ending in A or T vastly outnumber the synonymous codons ending in G or C. No in-frame ATG start codon is present in a number of genes in *E. huxleyi* cpDNA. Protein alignments suggest that GTG may serve as the codon for translation initiation in the genes *psbE*, *psbZ*, *rbcR*, *rpl33*, *rps3*, and *ycf27*. This has been reported for the plastid genomes of several different organisms, such as

Table III.3. Codon usage in the chloroplast genome of *E. huxleyi*<sup>a</sup>

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	840	1.33	Ser	UCU	344	1.2
	UUC	419	0.67		UCC	118	0.41
Leu	UUA	1458	2.99	Pro	UCA	496	1.74
	UUG	211	0.43		UCG	186	0.65
	CUU	569	1.17		CCU	320	1.29
	CUC	106	0.22		CCC	75	0.3
	CUA	459	0.94		CCA	463	1.87
	CUG	125	0.26		CCG	133	0.54
Ile	AUU	1329	2.04	Thr	ACU	452	1.2
	AUC	364	0.56		ACC	123	0.33
	AUA	264	0.4		ACA	710	1.88
Met	AUG	525	1	Ala	ACG	226	0.6
Val	GUU	915	1.97		GCU	633	1.42
	GUC	137	0.29		GCC	211	0.47
	GUA	594	1.28		GCA	699	1.56
	GUG	213	0.46		GCG	246	0.55
Tyr	UAU	516	1.27	Cys	UGU	168	1.53
	UAC	295	0.73		UGC	51	0.47
Ter	UAA	159	1.92	Ter	UGA	13	0.16
	UAG	76	0.92		UGG	332	1
His	CAU	190	0.92	Arg	CGU	557	2.57
	CAC	224	1.08		CGC	124	0.57
Gln	CAA	853	1.56		CGA	271	1.25
	CAG	243	0.44		CGG	78	0.36
Asn	AAU	817	1.32	Ser	AGU	408	1.43
	AAC	422	0.68		AGC	162	0.57
Lys	AAA	1270	1.49	Arg	AGA	223	1.03
	AAG	429	0.51		AGG	45	0.21
Asp	GAU	816	1.55	Gly	GGU	928	2.01
	GAC	238	0.45		GGC	204	0.44
Glu	GAA	1131	1.51		GGA	511	1.11
	GAG	365	0.49		GGG	203	0.44

<sup>a</sup> Codons are indicated in upper case letters, amino acid residues are in three-letter code; termination codons (UAA, UAG and UGA) are indicated by ter. The total number of individual codons (N) and the relative synonymous codon usage (RSCU) are shown.

Table III.4. List of genes encoded in the plastid genome of *E. huxleyi*

Classification	N	Genes							
Genetic system	(43)								
RNA Polymerase	4	<i>rpoA</i>	<i>rpoB</i>	<i>rpoC1</i>	<i>rpoC2</i>				
Transcription factors	2	<i>rbcR</i>	<i>ycf27</i>	<i>(ompR)</i>					
Translation	1	<i>tufA</i>							
Ribosomal proteins									
large subunits	16	<i>rpl2</i>	<i>rpl3</i>	<i>rpl5</i>	<i>rpl6</i>	<i>rpl14</i>	<i>rpl16</i>	<i>rpl19</i>	<i>rpl20</i>
		<i>rpl21</i>	<i>rpl22</i>	<i>rpl23</i>	<i>rpl27</i>	<i>rpl31</i>	<i>rpl33</i>	<i>rpl34</i>	<i>rpl36</i>
small subunits	17	<i>rps2</i>	<i>rps3</i>	<i>rps4</i>	<i>rps5</i>	<i>rps6</i>	<i>rps7</i>	<i>rps8</i>	<i>rps9</i>
		<i>rps10</i>	<i>rps11</i>	<i>rps12</i>	<i>rps13</i>	<i>rps14</i>	<i>rps16</i>	<i>rps17</i>	<i>rps18</i>
		<i>rps19</i>							
Protein quality control	3	<i>clpC</i>	<i>dnaK</i>	<i>groEl</i>					
Photosystems	(35)								
Photosystem I	11	<i>psaA</i>	<i>psaB</i>	<i>psaC</i>	<i>psaD</i>	<i>psaF</i>	<i>psaI</i>	<i>psaJ</i>	<i>psaL</i>
		<i>psaM</i>	<i>ycf3</i>	<i>ycf4</i>					
Photosystem II	15	<i>psbA</i>	<i>psbB</i>	<i>psbC</i>	<i>psbD</i>	<i>psbE</i>	<i>psbF</i>	<i>psbH</i>	<i>psbI</i>
		<i>psbJ</i>	<i>psbK</i>	<i>psbL</i>	<i>psbN</i>	<i>psbT</i>	<i>psbV</i>	<i>psbZ</i>	
Cytochrome complex	9	<i>petA</i>	<i>petB</i>	<i>petD</i>	<i>petG</i>	<i>petL</i>	<i>petM</i>	<i>petN</i>	<i>ccsA</i>
ATP synthesis	(8)								
ATP synthase	8	<i>atpA</i>	<i>atpB</i>	<i>atpD</i>	<i>atpE</i>	<i>atpF</i>	<i>atpG</i>	<i>atpH</i>	<i>atpI</i>
Metabolism	(5)								
Carbohydrates	3	<i>rbcL</i>	<i>rbcS</i>	<i>cbbX</i>					
Cofactors	2	<i>chlI</i>	<i>thiG</i>						
Cell processes	(2)								
Septum-site determination	1	<i>minD</i>							
Signal transduction	1	<i>dfr</i>							
Transport	(3)								
Transport	3	<i>secA</i>	<i>secY</i>	<i>sufB</i>					
Unknown	(23)								
Hypothetical proteins	14	<i>ycf12</i>	<i>ycf19</i>	<i>ycf20</i>	<i>ycf35</i>	<i>ycf39</i>	<i>ycf40</i>	<i>ycf43</i>	<i>ycf44</i>
		<i>ycf45</i>	<i>ycf46</i>	<i>ycf47</i>	<i>ycf55</i>	<i>ycf60</i>	<i>ycf65</i>	<i>ycf80</i>	
Unique ORFs	9	<i>orf32</i>	<i>orf33</i>	<i>orf35</i>	<i>orf36</i>	<i>orf40</i>	<i>orf42</i>	<i>orf46</i>	<i>orf71</i>
		<i>orf132</i>							
RNA genes	(36)								
rRNAs	6	<i>rrn16</i>	<i>rrn23</i>	<i>rrn5</i>	<i>rrn16</i>	<i>rrn23</i>	<i>rrn5</i>		
tRNAs	30	<i>trnA</i>	<i>trnA</i>	<i>trnC</i>	<i>trnD</i>	<i>trnE</i>	<i>trnF</i>	<i>trnG</i>	<i>trnG</i>
		<i>trnH</i>	<i>trnI</i>	<i>trnI</i>	<i>trnK</i>	<i>trnL</i>	<i>trnL</i>	<i>trnL</i>	<i>trnM</i>
		<i>trnM</i>	<i>trnM</i>	<i>trnN</i>	<i>trnP</i>	<i>trnQ</i>	<i>trnR</i>	<i>trnR</i>	<i>trnR</i>
		<i>trnS</i>	<i>trnS</i>	<i>trnT</i>	<i>trnV</i>	<i>trnW</i>	<i>trnT</i>		

*C. caldarium* (Glöckner et al. 2000) and *O. sinensis* (Kowallik et al. 1995), although not referring to the same genes.

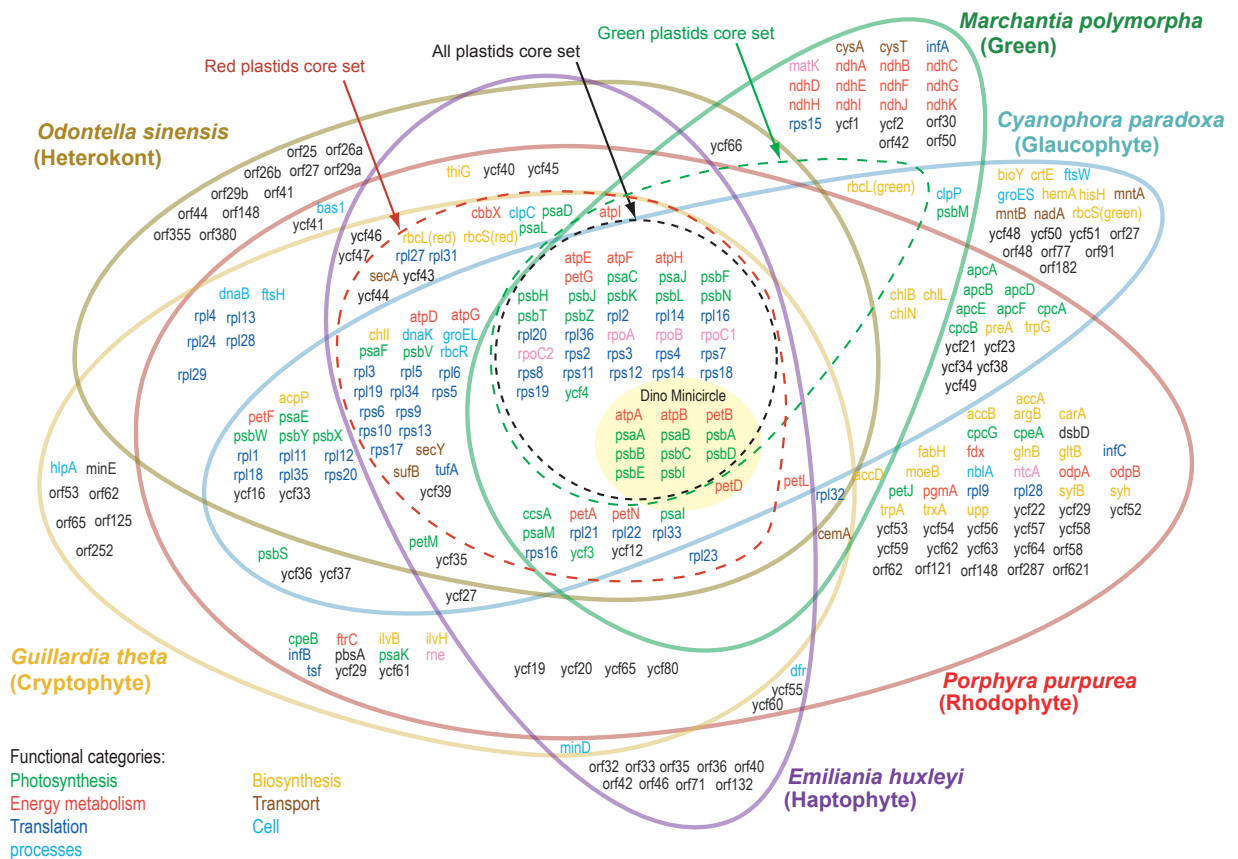
The plastid genome contains 30 tRNAs that are scattered throughout the entire genome, either singly or in groups, and all lack introns (Figure III.1, Table III.4). This set of plastid-encoded tRNAs is sufficient to decode the 62 sense codons that occur in protein-coding sequences, when taking into account wobble and the possible modifications of their anticodons (Crick 1966). The plastid genome of *E. huxleyi* contains three tRNA genes having the methionine anticodon CAU, which include the initiator and elongator methionine-accepting tRNAs.

### **Gene content in *E. huxleyi* cpDNA**

The plastid genome of *E. huxleyi* encodes 110 proteins. Table III.4 lists all the genes and ORFs by gene function present in *E. huxleyi* cpDNA. This represents the smallest number of genes in the red plastid genomes, which are characterized by containing a more comprehensive set of genes compared to the green algal and glaucophyte plastid genomes (Sugiura 1995) (see below). None of the protein coding genes contains introns. This is consistent with the other six red plastids, although introns are present in the plastid genomes of the euglenophyte *Euglena gracilis* (Hallick et al. 1993), the chlorophyte *Chlorella vulgaris* (Wakasugi et al. 1997), the flowering plants *Nicotiana tabacum* and *Oryza sativa* (Shimada and Sugiura 1991), and the liverwort *Marchantia polymorpha* (Shimada and Sugiura 1991). In addition to protein-coding genes, a total of 36 RNA genes are present in the *E. huxleyi* plastid genome, coding for 30 tRNAs and two copies of the genes encoding for the three



Figure III.2. Venn diagram comparing the protein-coding gene content of six plastid genomes. Core sets of genes from all plastids, from red plastids, and from green plastids were inferred from all 36 photosynthetic plastid genomes listed in Table III.1. Genes are colored depending on their functional category.



subunits of ribosomal RNA, namely small subunit (*rrs*), large subunit (*rrl*), and the 5S rRNA. I searched for unique open reading frames (ORFs) longer than 30 codons starting with ATG initiation codon. Nine ORFs were present that lack significant similarity to any entry in the public domain sequence databanks. The names of the ORFs correspond to the number of amino acids in the putative protein encoded by that gene.

### **Comparison of gene content and function among all plastid genomes**

Comparisons among all known plastid genomes from a wide range of photosynthetic eukaryotes have been made, and the gene content of six of those plastid genomes has been included in Figure III.2. Gene content of the plastid genomes of the three primary plastid lineages (red algae, green algae and glaucophytes), and the chl *c* containing algae (heterokonts, cryptophytes and haptophytes), are shown. Peridinin-containing dinoflagellates appear to have undergone an extreme organellar genome reduction and only a few genes thought to be in the chloroplast genome have been identified. I included the twelve distinct minicircle genes that have been described to date from different species of peridinin-containing dinoflagellates (Hiller 2001; Zhang et al. 2002; Laatsch et al. 2004; Nisbet et al. 2004).

Figure III.2 also shows the genes colored by functional category. Eight functional categories have been defined. “Photosynthesis” includes genes involved in light absorption, such as subunits of the photosystem I and II, and genes encoding phycobilisome proteins, among others. The functional category named “Energy metabolism” includes genes encoding proteins involved in the electron transport

chain, ATP synthesis, and pyruvate and acetyl-CoA metabolism. “Translation” refers to all genes related to protein synthesis, namely ribosomal proteins, initiation and elongation factors. “Transcription” includes RNA polymerases, and transcriptional regulators. The category named “Biosynthesis” holds genes involved in biosynthesis of amino acids, cofactors, fatty acids, chlorophyll, and carbohydrates. Genes encoding for proteins in charge of cell transport are listed under “Transport”. “Cell processes” is a diverse category including genes involved in cell division, septum site determining, protein folding, secretion pathway, detoxification, and degradation, among others. Hypothetical proteins (*ycf* genes) are listed as “Hypothetical”.

Excluding dinoflagellates, a core set of 45 genes (all plastids core set, Figure III.2) is present in all 36 photosynthetic plastid genomes sequenced to date (Table III.1). This core set of genes has been noted before (Martin et al. 1998; Martin et al. 2002; Grzebyk et al. 2003; Bachvaroff et al. 2004), and its composition remains fairly constant despite the addition of several new genomes. These genes represent several functional categories, but the majority are involved in the processes of photosynthesis, electron transfer, and protein synthesis. It was suggested that the expression of genes involved in key roles in electron transport and energy coupling is regulated tightly through redox potentials generated by the same electron transfer (Allen and Raven 1996; Pfannschmidt et al. 1999; Race et al. 1999). Therefore, the genes encoding these proteins are required to be in the organelle to be able to respond rapidly to maintain redox balance (Allen and Raven 1996; Pfannschmidt et al. 1999; Race et al. 1999), and this could partially explain the retention of genes in the

organelles. The type of genes retained in all plastid genomes (“all plastid core set”) is consistent with this hypothesis.

A more comprehensive set of genes, 93, is retained in all known photosynthetic red plastid genomes (“red plastids core set”, Figure III.2), including those of red algae, cryptophytes, heterokonts, and haptophytes, but excluding the extremely reduced plastid genome of dinoflagellates. Most of these genes are involved in photosynthesis, electron transport, and translation, among other cellular processes. This similarity in gene content among the red plastids supports the hypothesis that haptophyte, cryptophyte, and heterokont plastids are related, and were obtained from the red lineage. The number of genes retained in the red plastid genomes is higher than that of all photosynthetic “green” plastid genomes, including plants, green algae, and euglenophytes. The “green plastids core set” represents 47 genes (Figure III.2), and 50 (including *clpP*, *petA* and *ycf3*) when the secondary plastids of euglenophytes are excluded. This could be a consequence of the higher number of green plastid genomes that have been sequenced compared to the number of red plastid genomes (32 vs. 7, Table III.1).

Overall, the plastid genomes of the green lineage contain more genes involved in energy metabolism (NADH plastoquinone oxidoreductase subunits, in particular) compared to the red plastid genomes. In contrast, plastid genomes of the red lineage contain more genes related to translation (mostly ribosomal proteins), biosynthesis of amino acids, fatty acids, and pigments, and a variety of cell processes. The higher number of genes present in the red plastid genomes (Sugiura 1995) may account for the diverse type of processes carried out by plastid genes in these lineages.

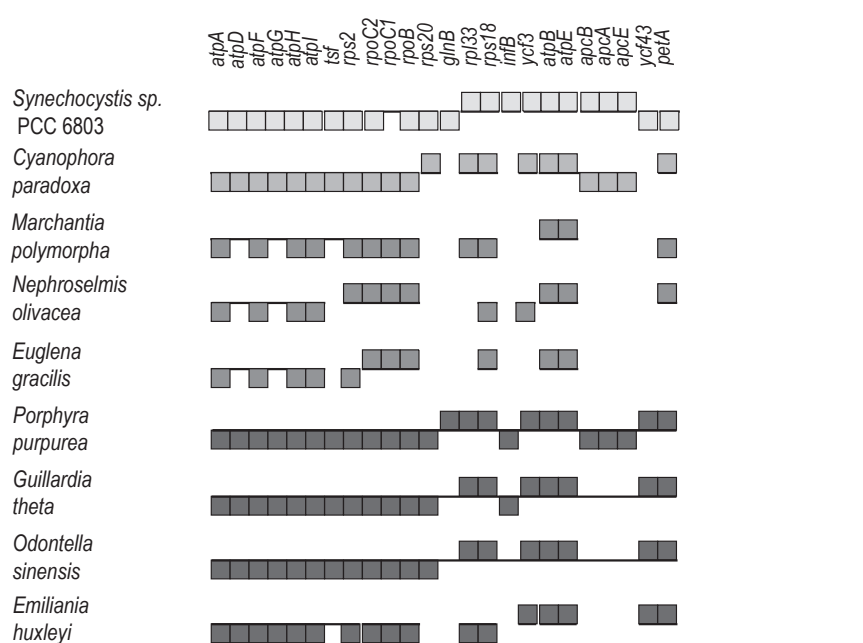
## Cluster analysis

Comparison of gene order in sequenced plastid genomes provides a powerful tool for inferring events in plastid evolution (Stoebe and Kowallik 1999).

Furthermore, it provides clear insights into evolutionary affiliation among algae. I examined the organization of genes in a number of plastid genomes by gene-cluster analysis. Plastid genomes of the chl *c* containing algae underwent major rearrangements since the acquisition of the secondary plastids from the red lineage. Only two partially conserved gene-clusters were found (Figure III.3).

The arrangement of the genes in the ribosomal operon observed in the cyanobacterium *Synechocystis* sp. PCC 6803 (Kaneko et al. 1996) is conserved in all plastid genomes so far sequenced, including some modifications, most of which are gene losses or transfers to the nuclear genome (Figure III.3, left side). This arrangement supports the hypothesis that all plastids are monophyletic (Delwiche et al. 1995; Stoebe and Kowallik 1999), and are originally derived from endosymbiotic cyanobacteria. The genes in the ribosomal operon are located on the same strand, and they were shown to be transcribed all together in the cryptophyte *G. theta* (Wang et al. 1997). A five-gene operon, including *rps12*, *rps7*, *fus*, *tufA* and *rps10*, displays a distinct arrangement in all red plastids. This operon is appended to the *rpl31* gene (3' end of the ribosomal operon), and it is located on the same strand (Stoebe and Kowallik 1999). It was proposed that the translocation of the five-gene cluster occurred early after the divergence of the rhodophyte lineage from the green lineage

Figure 1: Phylogenetic tree and presence of rRNA genes in various algae. The tree shows relationships between *Synechocystis* sp. PCC 6803, *Cyanophora paradoxa*, *Marchantia polymorpha*, *Nephroselmis olivacea*, *Euglena gracilis*, *Porphyra purpurea*, *Guillardia theta*, *Odontella sinensis*, and *Emiliana huxleyi*. The rRNA genes are indicated by black boxes above the tree. The genes are: rpl3, rpl4, rpl23, rpl2, rps19, rpl22, rps3, rpl16, rpl29, rps17, rpl14, rpl24, rpl5, rps8, rpl6, rpl18, rps5, rpl15, secY, adk, infA, rpl36, rps13, rps11, rps4, rpl3, rps9, rpl31, rps12, rps7, rplA, and rps10.



and glaucophytes (Ohta et al. 1997). This arrangement supports the relationship of chlorophyll c containing plastids with red algal plastids.

Another conserved region in the plastid genomes includes genes involved in diverse functions in the cell, such as ATP synthesis (*atpA-I*), RNA polymerization (*rpoB-C2*), electron transport (*petA*), etc. (Figure III.3, right side). These genes are encoded on different DNA strands. Many of the genes from this group are scattered throughout the chloroplast genome in the green lineage and glaucophytes. The order of these genes, including some gene losses, is conserved in the red algal plastids and *G. theta* and *O. sinensis* cpDNAs, suggesting again how closely related these plastids are (Stoebe and Kowallik 1999). In contrast, genes from this group are distributed over six independent regions in the *E. huxleyi* cpDNA, probably due to extensive genome rearrangements after acquisition of the plastid.

### **Parsimony analysis of presence and absence of genes**

To examine the relationships among plastid genomes, I analyzed 261 protein-coding genes that occurred among 18 plastid genomes from unrelated hosts (Table III.2). I constructed a dataset based on presence and absence of plastid genes, under the assumption that all of the genes were present in the cyanobacterial ancestor (although this assumption is probably violated at least for *rbcL*, (Delwiche and Palmer 1996). These data were analyzed with Camin-Sokal parsimony (Figure III.4), which assumes that genes lost from a plastid genome cannot be regained. Analyses were also performed with unweighted Fitch parsimony (Figure III.5), but showed a topology inconsistent in many ways with other phylogenetic information (Yoon et al.

Figure III.4. Single most parsimonious tree based on presence and absence of 261 plastid genes in the plastid genomes of 18 eukaryotes and the genome of the cyanobacterium *Synechocystis* sp. The tree was found by branch and bound search in PAUP\*4b10 using Camin-Sokal parsimony. Numbers above branches are bootstrap support values obtained from 1,000 replicates. Branch lengths are proportional to the number of character-state changes, and are shown numerically by values below the branches.

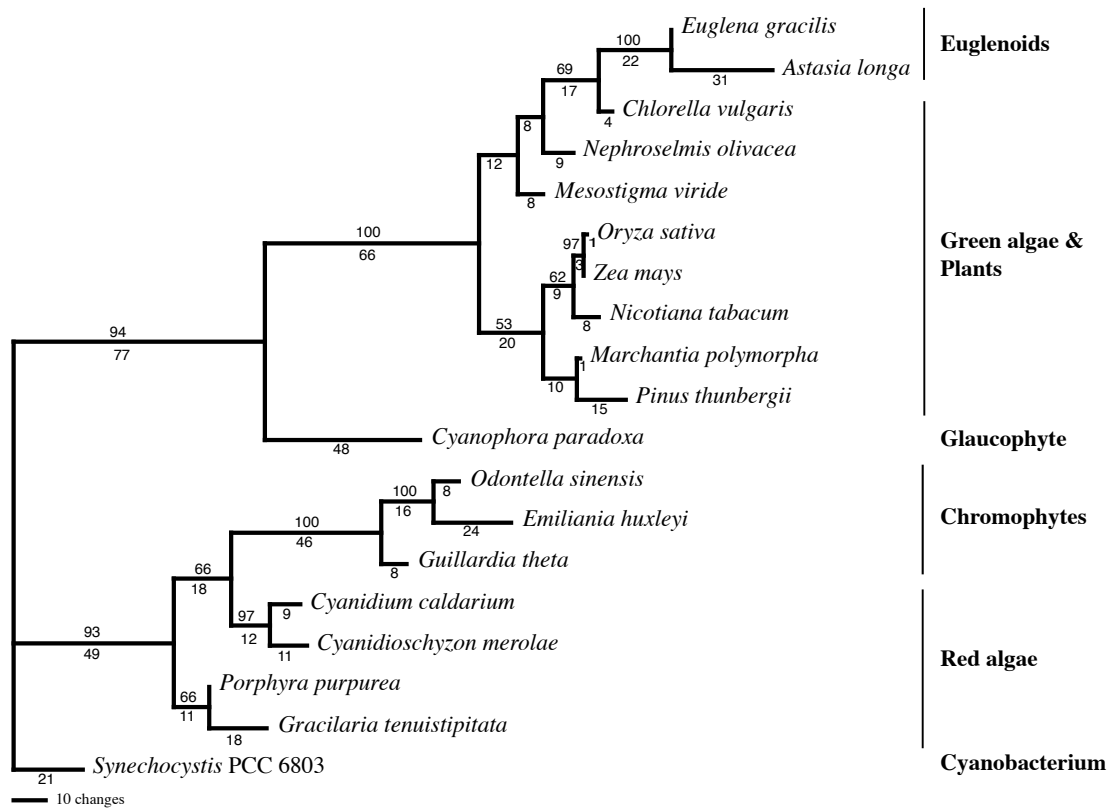
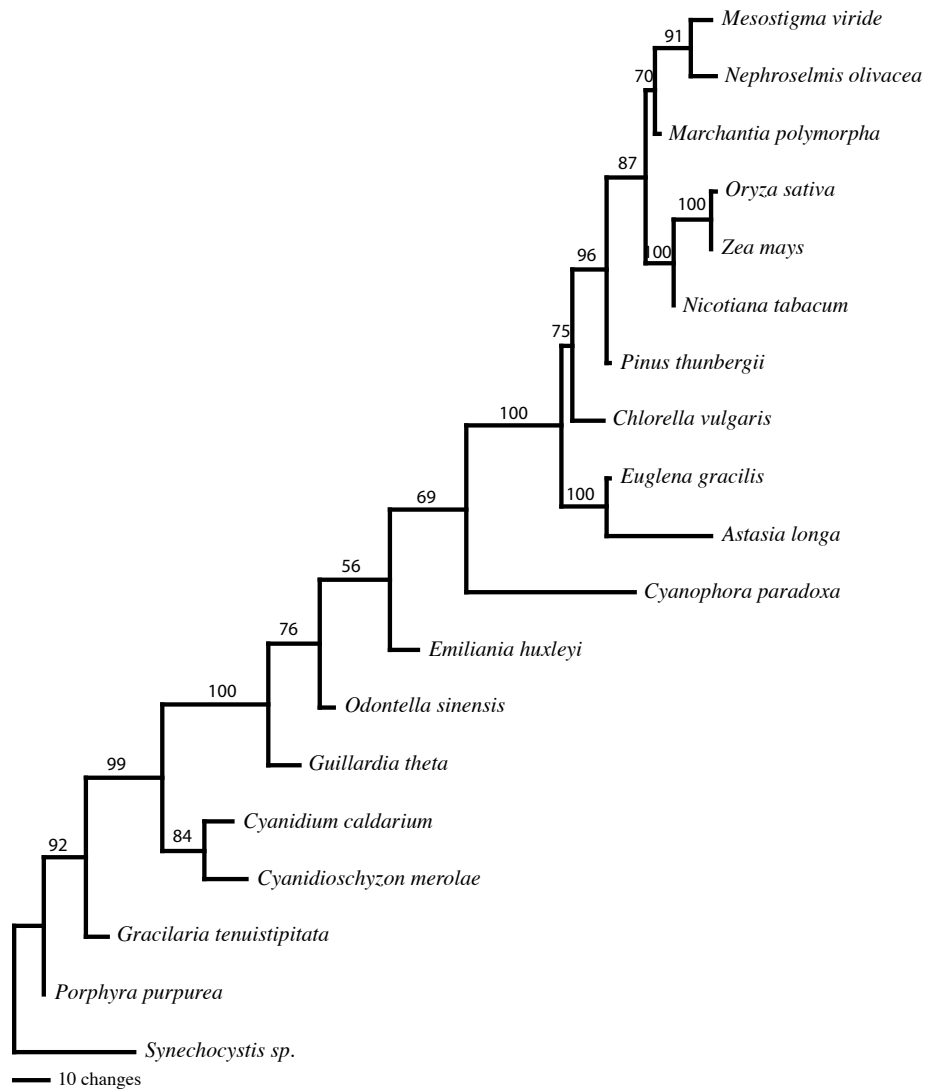




Figure III.5. Single most parsimonious tree based on presence and absence of 261 plastid genes (Table III.2) found by branch and bound search in PAUP\* using Fitch parsimony. Numbers above branches are bootstrap values. Branch lengths are proportional to the number of character-state changes.



2002b; Patron et al. 2004; Yoon et al. 2004). I attributed this discrepancy to the apparent independent loss of several genes (Martin et al. 2002), as seen in Figure III.2, and on this basis abandoned further consideration of the Fitch tree. The most parsimonious tree under Camin-Sokal parsimony was 620 steps long, with a consistency index of 0.389 and retention index of 0.847. Most of the genes from the cyanobacterial endosymbiont are thought to have been lost or transferred to the host nuclear genome soon after primary endosymbiosis. Two major clades were recovered and correspond to the red and green plastid lineages, including in each group the secondary plastids derived from them. The glaucophyte *C. paradoxa* lies between these two groups, but on a branch uniting it with the green lineage. The phylogenetic relationships of the glaucophyte plastid are still unknown, since its position varies in different studies and it is not strongly supported (Martin et al. 1998; Grzebyk et al. 2003; Chu et al. 2004). In general, green plastids encode fewer genes than red plastids, but the plastid genome of *E. huxleyi* encodes for a total number of genes that is comparable to the green plastids, with a markedly different gene content (see above). The phylogenetic relationships observed in the tree agree with relationships recovered from analyses based on sequence analysis, using single genes or a concatenation of a number of plastid genes (Yoon et al. 2002b; Harper and Keeling 2003; Patron et al. 2004; Yoon et al. 2004). In the present analysis the chromophyte plastids, those of the haptophyte, the heterokont and the cryptophyte, form a monophyletic group embedded in the red lineage. This is consistent with the origin of the chromophyte plastids from the red algal lineage, as shown in previous studies (Yoon et al. 2002b; Grzebyk et al. 2003; Nozaki et al. 2003a). Even though the

monophyly of the chl c containing plastids is supported, these data do not test or support the monophyly of the host cells, which continues to be partially understood at best.

## **Chapter IV - Phylogenetic Signal vs. Noise in Plastid**

### **Genomic Data and the Evolution of Chlorophyll c**

#### **Containing Plastids**

##### ***Abstract***

Photosynthetic eukaryotes contain primary, secondary, or tertiary plastids, depending on the source of the organelle (a cyanobacterium or a photosynthetic eukaryote). Plastid evolutionary history is quite complex and has been previously analyzed using plastid genomic data, although with poor taxon sampling, or a reduced dataset, or both. Although those studies could not fully resolve plastid relationships, several analytical problems, such as covarion evolution, and compositional bias were recognized. Here, I present an analysis of a multi-gene dataset based on 62 plastid-associated genes of 15 taxa representing the major plastid lineages. In an attempt to distinguish phylogenetic signal from noise, the data were analyzed using a wide range of phylogenetic methods (maximum parsimony, distance, and maximum likelihood analyses). Approximately unbiased tests, as well as homogeneity tests, were used to assess confidence on the results. The data suggest that primary plastids from glaucophytes branched before the divergence of green and red algal primary plastids. The chl c containing plastids are monophyletic and acquired their plastids from the red algae after the emergence of the Cyanidiales. The relationships among chl c containing plastids are hard to resolve. The data indicate that cryptophyte plastids are basal to the chl c plastid clade, and that the haptophyte and peridinin-containing

dinoflagellate plastids are sister taxa. However, the number of secondary endosymbioses that took place in the evolution of chl c containing algae and the relationships among the host cells remains uncertain. At least two hypotheses of host cell evolution are congruent with the plastid tree presented here: the chromalveolate and serial hypotheses.

## ***Introduction***

Plastids have a complex evolutionary history that includes primary, secondary, and tertiary endosymbiotic events involving cyanobacteria and several eukaryotic lineages (Delwiche 1999; Keeling 2004). Studies of molecular data have partially clarified the pattern of plastid transfer and acquisition; complete plastid genomes provide support for some clades, although resolution of other phylogenetic questions is less clear, and different studies have at times reached conflicting conclusions. In the present study, I aim to analyze a large plastid gene dataset to test the evolution of plastids, with an emphasis on the red-lineage plastids. To understand the reasons for conflict in previous studies, I applied a number of tests that assess possible deviations of the data from the assumptions and perform a wide range of phylogenetic analyses.

Oxygenic photosynthesis arose once in evolution in cyanobacteria and was later acquired by unrelated eukaryotic lineages in several independent evolutionary events (Delwiche 1999). A primary endosymbiotic event is one in which a heterotrophic eukaryote engulfed a cyanobacterium and retained it as a permanent endosymbiont. Following the establishment of endosymbiosis, the cyanobacterium lost many of its genes, many others were transferred to the host nucleus, and ca. 100-

200 genes, depending upon the lineage, were retained as the plastid genome. Primary plastids are surrounded by two membranes that are thought to correspond to the inner and outer membranes of the cyanobacterial endosymbiont, and are found in at least three algal lineages: red algae, green algae (including land plants), and glaucophytes (Keeling 2004). Secondary endosymbiotic events are those where a heterotrophic eukaryote acquired photosynthesis by engulfing a photosynthetic eukaryote (green or red alga). In some cases, organisms with secondary plastids have themselves become endosymbionts. Four photosynthetic lineages, namely haptophytes, dinoflagellates, cryptophytes, and heterokonts, contain secondary plastids surrounded by 3-4 membranes, with chlorophyll c as a main photosynthetic pigment. These plastids are thought to be derived ultimately from the red algal lineage (Delwiche and Palmer 1997; Durnford et al. 1999; Yoon et al. 2002b; Bachvaroff et al. 2005). Even though not all the members of these four lineages are photosynthetic or contain a plastid, I will refer to them as the “chl c containing algae”. The plastids of the red algae (herein “red algal plastids”) together with the chl c containing plastids derived from them are here collectively called “red-lineage plastids”. The number of endosymbiotic events giving rise to the plastids of the chl c containing algae and the relationships among them has not yet been determined.

Relationships among plastid lineages have proven difficult to resolve with either DNA or protein sequence data. A number of phylogenies have been published on this topic with conflicting results; in many cases the taxon sampling was low, or analyses were based on a small number of genes, or both (Martin et al. 1998; Fast et al. 2001; Ishida and Green 2002; Martin et al. 2002; Yoon et al. 2002b; Bachvaroff et

al. 2005; Yoon et al. 2005). In addition to limitations imposed by data availability (i.e., gene and taxon sampling), potential shortcomings of the currently available cpDNA data include analytical problems due to covarion or/and heterogeneous evolution (Lockhart et al. 1998; Lockhart et al. 1999; Phillips et al. 2004; Ane et al. 2005). The central objective here is to analyze the available plastid data and study possible reasons for conflict. Many previous studies have applied only a narrow range of phylogenetic methods, and in at least some cases the analytical method seems to have influenced the conclusions. Recently, more data have become available, including the complete plastid genome of a haptophyte (Sanchez-Puerta et al. 2005) and EST projects on dinoflagellates (Bachvaroff et al. 2004; Hackett et al. 2004), that allowed us to perform phylogenetic analyses including all four lineages of chl c containing algae based on a large number of characters. In the present study, I examined a dataset fully representing the red-lineage plastids with as many genes as is practical, and applied a range of phylogenetic methods, as well as compositional homogeneity tests and approximately unbiased (AU) tests, to study the interaction between dataset composition and analytical method in terms of phylogenetic conclusions.

This study addresses four key issues of plastid evolution and hypotheses of host cell evolution: the position of the glaucophyte plastid, monophyly or paraphyly of red algal plastids, monophyly of the chl c containing plastids, and relationships among the chl c plastids. I analyzed a 62-gene dataset using a variety of phylogenetic methods including nucleotide and amino acid based analyses, as well as AU tests on concatenated data.

Table IV.1. Taxonomy of photosynthetic eukaryotes and taxon sampling for this study.

Division/Phylum	Genera included in this study
Chlorophyta + Streptophyta	<i>Arabidopsis</i> , <i>Chaetosphaeridium</i> , <i>Mesostigma</i> , <i>Nephroselmis</i> <sup>a</sup>
Rhodophyta	<i>Cyanidioschyzon</i> , <i>Cyanidium</i> , <i>Gracilaria</i> , <i>Porphyra</i> <sup>b</sup>
Glaucophyta	<i>Cyanophora</i> <sup>b</sup>
Cryptophyta	<i>Guillardia</i> <sup>b</sup>
Heterokontophyta (Stramenopiles)	<i>Odontella</i> <sup>b</sup>
Dinophyta (Pyrrophyta)	<i>Alexandrium</i> , <i>Amphidinium</i> , <i>Lingulodinium</i> <sup>b</sup>
Haptophyta (Prymnesiophyta)	<i>Emiliania</i> <sup>b</sup>
Euglenophyta	none <sup>c</sup>
Chlorarachniophyta	none <sup>c</sup>

<sup>a</sup> This is a subset of the complete plastid genomes publicly available for this group.

<sup>b</sup> Taxon sampling for these lineages include all current, publicly available data.

<sup>c</sup> Members of these groups acquired their plastid from the green algal lineage, and thus, they are out of the scope of this study.



## ***Materials and Methods***

### **Sequence acquisition and alignment**

Sequences were downloaded from GenBank database at the National Center for Biotechnology Information (NCBI), imported into MacClade 4.0 (Maddison and Maddison 2000), and aligned by eye. Recently acquired data from the plastid genome of the haptophyte *Emiliana huxleyi* were included (Sanchez-Puerta et al. 2005). Peridinin-containing dinoflagellate sequences are from *Amphidinium*, *Alexandrium* or *Lingulodinium*, depending on the availability. Table IV.1 shows the major lineages of photosynthetic eukaryotes and the taxon sampling used for this study. Table IV.2 lists the GenBank accession numbers of all the sequences included in the phylogenetic analyses.

Datasets including or excluding the dinoflagellate plastid were analyzed. The reason for this is that plastid-associated genes in dinoflagellates show a higher rate of evolution than those in other photosynthetic eukaryotes (Zhang et al. 2000; Bachvaroff et al. 2006). Different relative rates in different lineages (condition known as covarion evolution) have been implicated in incorrect phylogenetic inference due to long-branch attraction artifact (Inagaki et al. 2004b).

Table IV.2. Published sequences used in the phylogenetic analyses

Taxon name	GenBank accession number
<i>Arabidopsis thaliana</i>	NC_000932
<i>Alexandrium tamarense</i>	CF948052, CK432619, CK432854, CK433015, CK433113, CK433335, CK784212, CK784696, CK785316, CV554053, CV554882, CX769352
<i>Amphidinium operculatum</i>	AJ311628-AJ311632; AJ250262-AJ250266, AJ582639, CF065976, CF066904, CF067846, CF065182, CF065427, CF064650, CF067332, CF067650, CF067393, CF067587, CF064857, CF066016, CF067549
<i>Chaetosphaeridium globosum</i>	NC_004115
<i>Cyanidium caldarium</i>	NC_001840
<i>Cyanidioschyzon merolae</i>	NC_004799
<i>Cyanophora paradoxa</i>	NC_001675
<i>Emiliana huxleyi</i>	NC_007288
<i>Guillardia theta</i>	NC_000926
<i>Mesostigma viride</i>	NC_002186
<i>Nephroselmis olivacea</i>	NC_000927
<i>Nostoc sp.</i> PCC7120	NC_003272
<i>Odontella sinensis</i>	NC_001713
<i>Porphyra purpurea</i>	NC_000925
<i>Synechocystis sp.</i> PCC6803	NC_000911
<i>Lingulodinium polyedrum</i>	BP742897, BP743452

## Phylogenetic analyses

Single gene and concatenated nucleotide datasets were analyzed using PAUP\*4b10 (Swofford et al. 2002). Third codon positions were excluded from all analyses. For the maximum likelihood (ML) analyses, first, a Fitch-Margoliash tree was constructed using LogDet distances. Then, parameters for the ML were estimated from the distance tree. In the likelihood analysis, the General Time Reversible model with Invariant site and gamma correction was used (GTR + I + $\Gamma$ ) with four rate categories. The model of substitution was selected using hierarchical likelihood ratio tests from Modeltest v.3.5 (Posada and Crandall 1998). For the distance analyses, a minimum evolution (ME) tree using LogDet distances was constructed. Bootstrap analyses were performed using three random additions with nearest neighbor interchange.

Single-gene and concatenated datasets based on amino acids were analyzed with TreePuzzle 5.2 (Schmidt et al. 2002), under the JTT model of amino acid substitution, with eight rate categories and invariant sites estimated from the dataset. The concatenated analyses were also analyzed using ProML (PHYLP 3.63) under the JTT model of amino acid substitution with rate heterogeneity including invariant sites and gamma-distributed rates in eight categories. An estimation of the parameters was obtained from the analyses using TreePuzzle. For bootstrap analyses, 100 datasets were created using the SEQBOOT program in the PHYLIP package. The consensus bootstrap tree was obtained with the CONSENSE program of PHYLIP. In addition, maximum parsimony (MP) analyses based on amino acid and nucleotide data (only first and second positions) were performed using PAUP\*. To assess compositional

homogeneity of the data, the amino acid and nucleotide frequency distribution of each taxon was compared to the one estimated by the maximum likelihood model using TreePuzzle, and the significance of the differences was evaluated with a chi square test (Schmidt et al. 2002).

### **Approximately Unbiased (AU) tests**

The CONSEL package (Schimodaira and Hasegawa 2001; Schimodaira 2002) was used to assess confidence in tree selection by calculating the AU p values for the different hypotheses tested: red-lineage plastid monophyly, red algal plastid monophyly, chl c plastids monophyly, a relationship of the glaucophyte plastid either with the red-lineage plastids, the green lineage plastids or both, and alternate affiliations of *Mesostigma* in the green plastid clade. Constrained trees were compared to the most likely unconstrained tree. For the nucleotide-based analyses, the most likely tree was found under ML using PAUP\* using same parameters as before, and the site likelihoods for each tree were used to calculate the AU p values. For amino acid-based analyses, the codeml program from the PAML 3.14 package (Yang 1997) was used to obtain the site likelihoods of the constrained trees.

## ***Results***

### **Individual gene analyses**

Sixty-two individual gene sequences were analyzed and the support values for a number of taxon bipartitions or specific clades from single gene trees are shown in

Table IV.3. Individual phylogenetic analyses based on plastid-associated genes, including dinoflagellates.

Gene name	Analyses	Number of sites	Support values for different clades a													Other relationships b
			1	2	3	4	5	6	7	8	9	10	11	12	13	
<i>atpA</i>	TreePuzzle	505	98	-	62	-	62	-	-	-	54	-	-	-	73	-
<i>atpA</i>	PAUP-no dino	1010	75	+	-	-	96	51	-	52	60	-	-	-	-	-
<i>atpA</i>	PAUP all	1010	81	-	78	-	79	+	+	+	-	-	-	+	-	-
<i>atpB</i>	TreePuzzle	476	97	-	-	-	78	-	-	-	-	-	-	92	-	-
<i>atpB</i>	PAUP-no dino	952	100	-	-	-	76	-	-	-	-	-	-	-	-	-
<i>atpB</i>	PAUP all	952	101	+	-	-	+	+	-	-	+	-	-	-	-	-
<i>atpE</i>	TreePuzzle	133	92	-	-	-	82	-	-	-	-	-	-	-	-	-
<i>atpE</i>	PAUP-no dino	266	74	-	-	-	81	-	-	-	-	-	-	-	-	-
<i>atpE</i>	PAUP all	266	66	-	-	-	73	-	-	-	-	-	-	-	-	-
<i>atpH</i>	TreePuzzle	81	56	-	-	+	-	-	-	83	83	-	-	-	-	-
<i>atpH</i>	PAUP-no dino	162	+	-	-	+	50	-	-	75	+	-	-	-	-	-
<i>atpH</i>	PAUP all	162	+	-	-	-	-	-	-	77	-	-	-	-	-	-
<i>atpI</i>	TreePuzzle	229	100	-	-	-	93	82	-	-	-	-	-	-	99	-
<i>atpI</i>	PAUP-no dino	458	99	-	-	-	60	98	-	-	-	-	-	-	-	-
<i>atpI</i>	PAUP all	458	+	-	-	-	71	96	-	-	-	-	-	-	100	-
<i>chlI</i>	TreePuzzle	335	52	+	-	-	78	-	-	-	-	-	-	-	-	-
<i>chlI</i>	PAUP-no dino	670	92	52	-	-	95	94	-	70	-	-	-	-	-	-
<i>chlI</i>	PAUP all	670	-	-	-	+	93	-	-	62	-	-	+	-	-	-
<i>clpC</i>	TreePuzzle	802	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>clpC</i>	PAUP-no dino	1604	-	-	-	-	-	-	-	-	-	-	51	-	-	-
<i>clpC</i>	PAUP all	1604	-	-	-	-	-	-	+	-	-	-	59	-	-	-
<i>dnaK</i>	TreePuzzle	541	64	-	-	55	98	+	-	-	-	-	83	-	-	-
<i>dnaK</i>	PAUP-no dino	1082	100	50	-	-	100	78	-	-	-	-	76	-	-	-
<i>dnaK</i>	PAUP all	1082	96	+	-	-	100	63	-	-	-	-	72	-	-	-
<i>petA</i>	TreePuzzle	290	77	-	-	51	-	-	-	-	64	-	-	-	-	-
<i>petA</i>	PAUP-no dino	580	100	57	-	-	85	+	-	-	-	-	-	-	-	-
<i>petA</i>	PAUP all	580	73	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>petB</i>	TreePuzzle	220	95	-	-	-	-	-	83	-	74	-	-	-	93	-
<i>petB</i>	PAUP-no dino	440	99	-	-	-	-	-	70	+	52	-	-	-	-	-
<i>petB</i>	PAUP all	440	90	-	-	-	-	-	+	-	+	-	-	-	+	-
<i>petD</i>	TreePuzzle	161	76	-	-	70	84	99	-	-	-	-	68	-	-	-
<i>petD</i>	PAUP-no dino	322	56	60	-	-	+	85	-	-	-	-	-	-	-	-
<i>petD</i>	PAUP all	322	+	+	-	-	-	78	-	+	-	+	-	-	-	-
<i>petG</i>	TreePuzzle	34	56	-	-	-	81	-	-	-	-	-	-	-	-	-
<i>petG</i>	PAUP-no dino	68	67	-	-	-	50	-	-	-	-	+	-	-	-	-
<i>petG</i>	PAUP all	68	66	-	-	-	+	+	-	-	-	+	-	-	-	-
<i>psaA</i>	TreePuzzle	755	73	-	-	65	64	62	-	-	-	-	92	-	-	-
<i>psaA</i>	PAUP-no dino	1510	100	85	-	-	97	85	75	-	-	-	66	-	-	-
<i>psaA</i>	PAUP all	1510	96	+	-	-	75	+	+	-	-	-	+	-	-	-
<i>psaB</i>	TreePuzzle	741	89	-	-	85	58	+	-	-	63	-	-	-	-	-
<i>psaB</i>	PAUP-no dino	1482	100	-	-	89	95	100	-	-	67	-	-	-	-	-
<i>psaB</i>	PAUP all	1482	70	-	-	56	96	65	-	-	-	-	-	-	-	-
<i>psaC</i>	TreePuzzle	83	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>psaC</i>	PAUP-no dino	166	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>psaC</i>	PAUP all	166	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>psaF</i>	TreePuzzle	156	86	51	-	-	-	-	-	-	-	-	-	50	-	-
<i>psaF</i>	PAUP-no dino	312	100	-	-	-	-	67	-	-	-	-	-	-	-	-
<i>psaF</i>	PAUP all	312	100	67	-	-	-	-	-	-	-	-	-	-	-	-
<i>psaJ</i>	TreePuzzle	35	-	-	-	-	76	-	-	-	-	-	-	-	-	-
<i>psaJ</i>	PAUP-no dino	70	+	-	-	-	+	+	-	-	-	-	+	-	-	-
<i>psaJ</i>	PAUP all	70	-	-	51	-	+	-	-	-	-	-	-	-	-	-
<i>psaL</i>	TreePuzzle	94	-	-	-	-	-	-	-	-	-	-	-	68	-	-
<i>psaL</i>	PAUP-no dino	188	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>psaL</i>	PAUP all	188	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>psbB</i>	TreePuzzle	505	74	-	-	61	100	87	-	-	-	-	50	-	-	-
<i>psbB</i>	PAUP-no dino	1010	100	-	-	+	100	100	-	71	-	-	80	-	-	-
<i>psbB</i>	PAUP all	1010	88	+	-	-	99	69	-	+	-	-	+	-	-	-
<i>psbC</i>	TreePuzzle	478	74	-	-	60	99	82	-	-	-	-	-	-	-	-
<i>psbC</i>	PAUP-no dino	956	100	-	-	62	100	100	+	72	-	-	97	-	-	-
<i>psbC</i>	PAUP all	956	96	-	-	76	91	65	-	-	-	-	+	-	-	-
<i>psbD</i>	TreePuzzle	339	98	-	94	-	100	70	-	-	-	-	-	-	-	-
<i>psbD</i>	PAUP-no dino	678	100	-	-	56	100	100	52	79	-	-	99	-	-	-
<i>psbD</i>	PAUP all	678	98	74	-	-	100	90	-	+	-	-	+	-	-	-
<i>psbE</i>	TreePuzzle	81	90	-	-	84	69	-	-	-	-	-	-	-	-	-
<i>psbE</i>	PAUP-no dino	162	85	-	-	-	50	-	+	-	+	-	-	-	-	-
<i>psbE</i>	PAUP all	162	95	-	-	+	+	-	-	-	+	-	-	-	-	-
<i>psbF</i>	TreePuzzle	37	68	-	-	+	57	70	-	-	-	-	-	80	-	-
<i>psbF</i>	PAUP-no dino	74	65	+	-	-	+	+	-	-	-	-	-	-	-	-
<i>psbF</i>	PAUP all	74	-	-	-	-	+	-	-	-	-	-	+	-	-	-
<i>psbH</i>	TreePuzzle	62	78	-	-	-	62	-	-	-	-	-	-	-	-	-
<i>psbH</i>	PAUP-no dino	124	63	66	-	-	90	-	-	-	-	-	-	-	-	-
<i>psbH</i>	PAUP all	124	52	-	-	-	65	-	-	-	-	-	-	-	-	-
<i>psbK</i>	TreePuzzle	41	54	-	-	-	-	-	-	-	-	-	-	-	78	-
<i>psbK</i>	PAUP-no dino	82	+	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>psbK</i>	PAUP all	82	-	-	-	-	+	-	-	-	-	-	-	-	-	-

<i>psbL</i>	TreePuzzle	39	94	-	-	-	-	-	-	-	-	-	-	-	79	-	-
<i>psbL</i>	PAUP-no dino	78	94	-	+	-	71	-	-	-	-	-	-	-	-	-	-
<i>psbL</i>	PAUP all	78	99	-	-	-	68	-	-	-	-	-	-	-	-	-	-
<i>psbN</i>	TreePuzzle	44	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>psbN</i>	PAUP-no dino	88	73	-	-	+	-	-	-	-	-	-	-	-	-	-	-
<i>psbN</i>	PAUP all	88	78	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>psbT</i>	TreePuzzle	30	96	-	-	-	56	-	-	-	-	-	-	-	-	-	-
<i>psbT</i>	PAUP-no dino	60	100	-	-	-	-	53	-	60	-	+	-	-	-	-	-
<i>psbT</i>	PAUP all	60	100	-	-	-	-	+	-	+	-	-	-	-	-	50	-
<i>rpl2</i>	TreePuzzle	275	77	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>rpl2</i>	PAUP-no dino	550	90	-	-	-	+	-	-	-	-	-	-	-	-	-	-
<i>rpl2</i>	PAUP all	550	82	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>rpl3</i>	TreePuzzle	193	51	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>rpl3</i>	PAUP-no dino	386	+	-	-	52	-	57	-	-	-	-	-	-	-	-	-
<i>rpl3</i>	PAUP all	386	56	-	-	+	-	-	-	-	-	-	-	-	-	-	-
<i>rpl5</i>	TreePuzzle	173	88	-	-	-	-	-	-	52	-	-	-	-	-	-	-
<i>rpl5</i>	PAUP-no dino	346	+	-	-	+	83	+	-	61	-	-	-	-	-	-	-
<i>rpl5</i>	PAUP all	346	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-
<i>rpl23</i>	TreePuzzle	85	73	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>rpl23</i>	PAUP-no dino	170	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>rpl23</i>	PAUP all	170	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>rpl33</i>	TreePuzzle	72	89	-	-	-	-	-	-	99	-	-	-	-	-	-	-
<i>rpl33</i>	PAUP-no dino	144	-	-	-	-	-	-	-	98	-	-	-	-	-	-	-
<i>rpl33</i>	PAUP all	144	+	-	-	-	-	-	-	91	-	-	-	-	-	-	-
<i>rps2</i>	TreePuzzle	226	86	-	-	84	-	-	-	70	-	-	-	-	-	-	-
<i>rps2</i>	PAUP-no dino	452	98	-	-	57	+	-	-	-	-	-	-	-	-	-	-
<i>rps2</i>	PAUP all	452	96	-	-	61	-	-	-	-	-	-	-	-	-	-	-
<i>rps9</i>	TreePuzzle	94	65	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>rps9</i>	PAUP-no dino	188	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>rps9</i>	PAUP all	188	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>secA</i>	TreePuzzle	442	93	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>secA</i>	PAUP-no dino	884	65	-	-	-	-	86	-	-	-	-	+	-	-	-	-
<i>secA</i>	PAUP all	884	+	-	-	-	-	-	-	-	-	-	+	-	-	-	-
<i>tufA</i>	TreePuzzle	411	81	-	-	-	-	-	-	70	-	-	-	-	-	+	+
<i>tufA</i>	PAUP-no dino	822	97	-	-	-	-	-	+	+	86	-	-	-	-	+	+
<i>tufA</i>	PAUP all	822	90	-	-	-	-	-	+	+	76	-	-	-	-	+	+
<i>ycf3</i>	TreePuzzle	166	-	-	-	76	95	63	-	-	-	-	-	-	-	-	-
<i>ycf3</i>	PAUP-no dino	332	-	-	-	57	100	71	+	-	-	-	50	-	-	-	-
<i>ycf3</i>	PAUP all	332	-	-	-	-	73	-	-	-	-	-	-	+	-	-	-

1- Cyanobacteria monophyly

2- Cyanophora + red plastids monophyly

3- Cyanophora + green plastids monophyly

4- Green + red plastids monophyly

5- Green plastids monophyly

6- Red plastids monophyly

7- Chl c containing plastids monophyly

8- Red algal plastids monophyly

9- Guillardia + Emiliania monophyly

10- Guillardia + Odontella monophyly

11- Emiliania + Odontella monophyly

12- Emiliania + dinoflagellate monophyly

13- Odontella + dinoflagellate monophyly

TreePuzzle- Protein analyses based on JTT model of substitution using the program TreePuzzle

PAUP- Nucleotide maximum likelihood analysis based on GTR+I+G4 model of evolution using PAUP\*

PAUP-no dino- ML analysis excluding the dinoflagellate

PAUP all- ML analysis including the dinoflagellate

a- Support values only when > 50. A dash (-) indicates that the clade is not present in the best tree

A plus (+) indicates that the clade is present in the best tree but the support value < 50.

b- Numbers correspond to support values for the relationships, when higher than 70

Figure IV.1. Maximum likelihood analyses based on individual plastid genes, under GTR+I+Γ4 using PAUP\*, including the peridinin-containing dinoflagellate (*Amphidinium* or *Alexandrium*). Numbers above branches are bootstrap support values, when > 60.

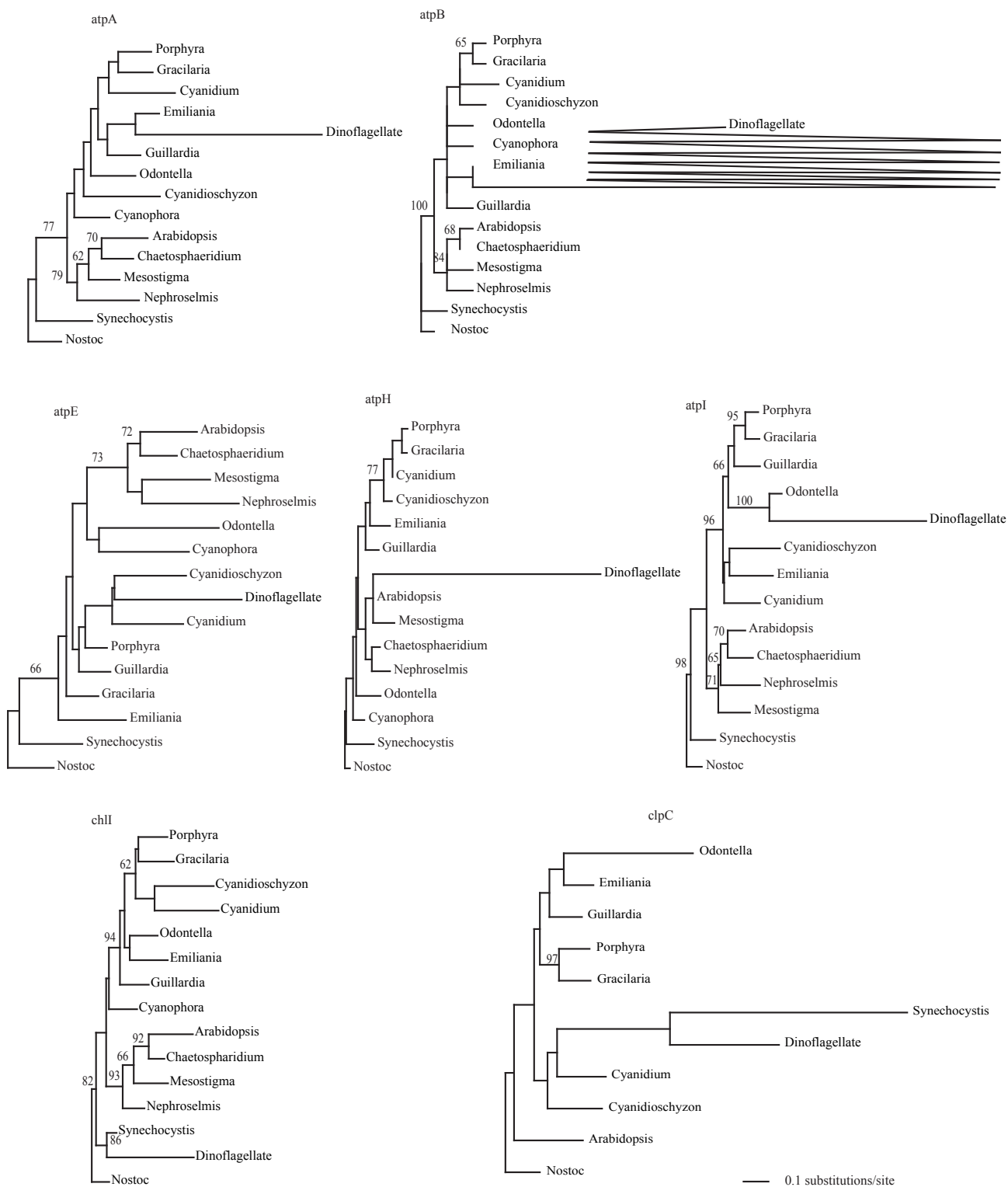


Figure IV.1 (con't). Maximum likelihood analyses based on individual plastid genes, under GTR+I+ $\Gamma$ 4 using PAUP\*, including the peridinin-containing dinoflagellate (*Amphidinium* or *Alexandrium*). Numbers above branches are bootstrap support values, when > 60.





Figure IV.1 (con't). Maximum likelihood analyses based on individual plastid genes, under GTR+I+ $\Gamma$ 4 using PAUP\*, including the peridinin-containing dinoflagellate (*Amphidinium* or *Alexandrium*). Numbers above branches are bootstrap support values, when > 60.

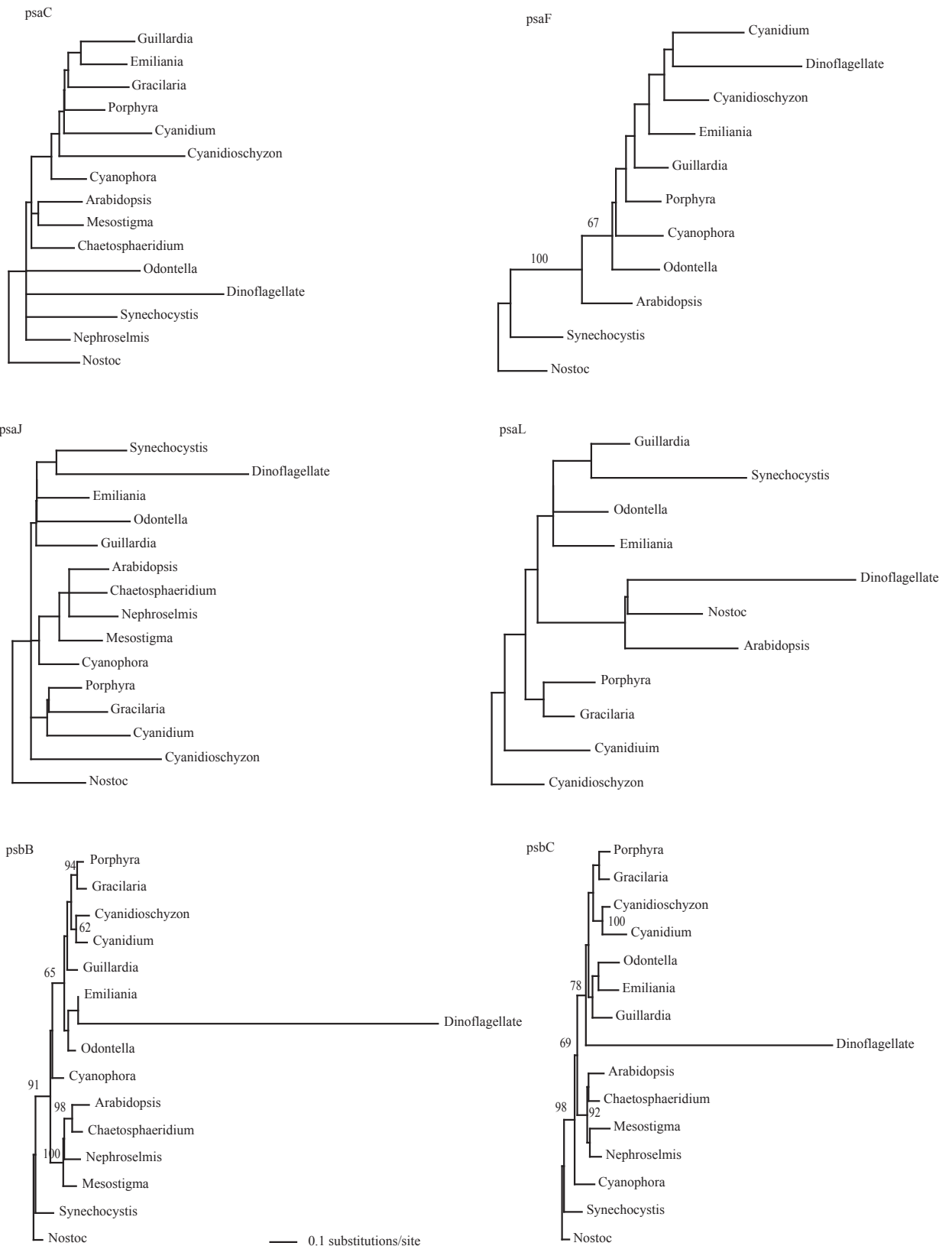


Figure IV.1 (con't). Maximum likelihood analyses based on individual plastid genes, under GTR+I+Γ4 using PAUP\*, including the peridinin-containing dinoflagellate (*Amphidinium* or *Alexandrium*). Numbers above branches are bootstrap support values, when > 60.

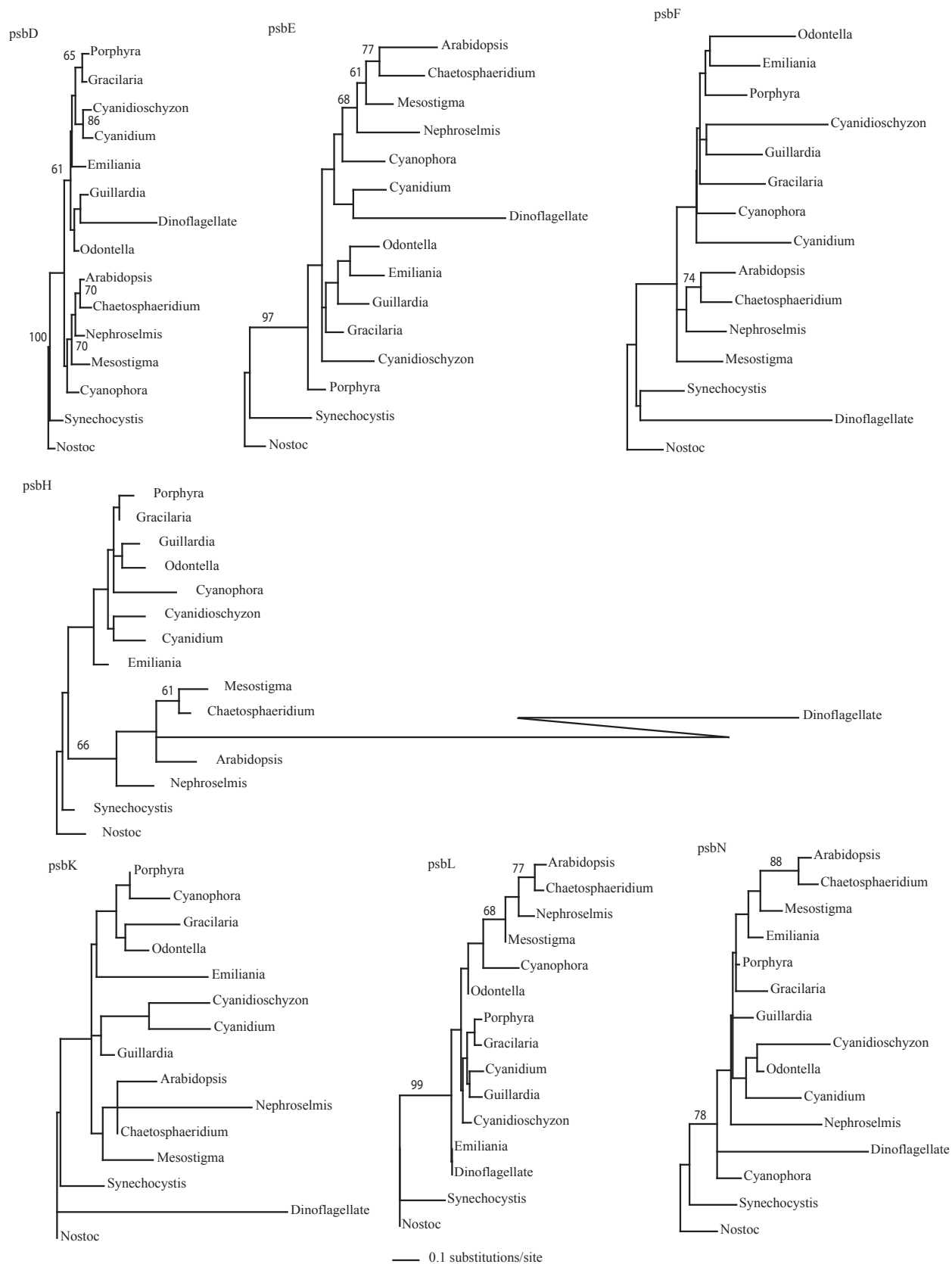


Figure IV.1 (con't). Maximum likelihood analyses based on individual plastid genes, under GTR+I+ $\Gamma$ 4 using PAUP\*, including the peridinin-containing dinoflagellate (*Amphidinium* or *Alexandrium*). Numbers above branches are bootstrap support values, when > 60.

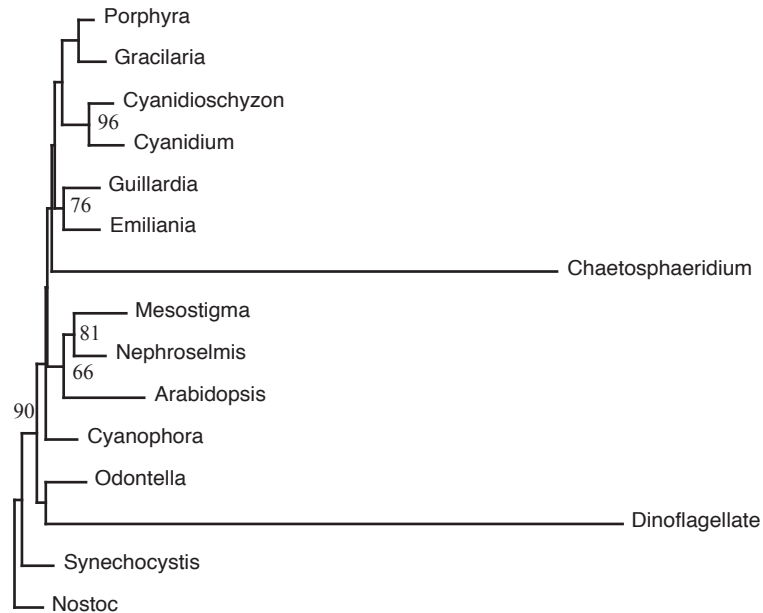


Figure IV.1 (con't). Maximum likelihood analyses based on individual plastid genes, under GTR+I+ $\Gamma$ 4 using PAUP\*, including the peridinin-containing dinoflagellate (*Amphidinium* or *Alexandrium*). Numbers above branches are bootstrap support values, when > 60.

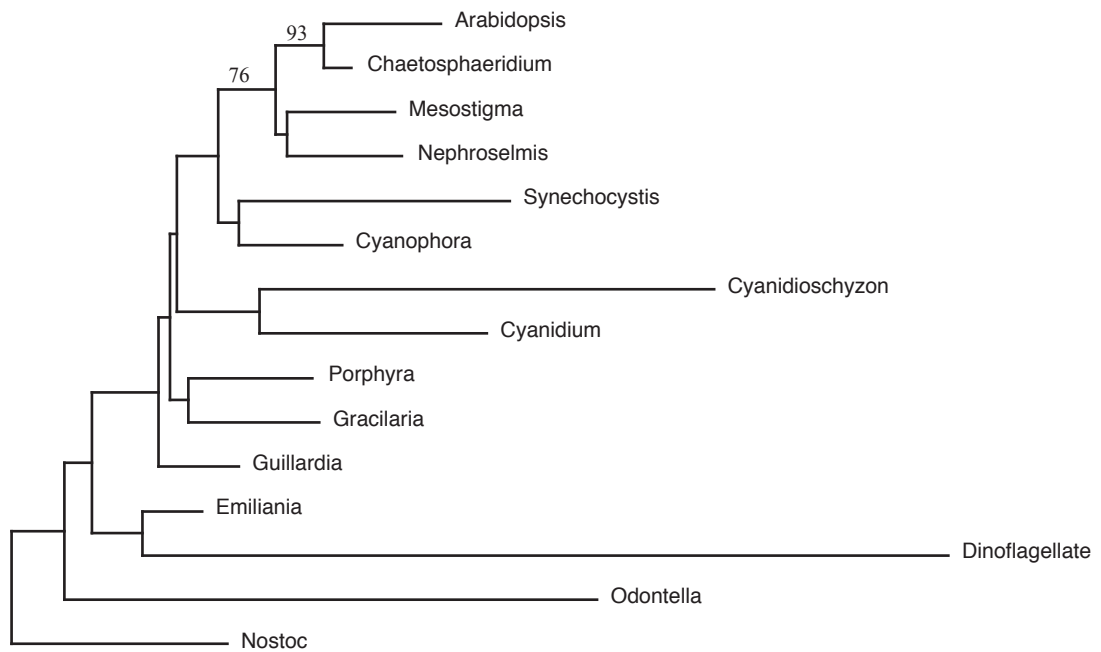


Figure IV.1 (con't). Maximum likelihood analyses based on individual plastid genes, under GTR+I+Γ4 using PAUP\*, including the peridinin-containing dinoflagellate (*Amphidinium* or *Alexandrium*). Numbers above branches are bootstrap support values, when > 60.

tufA



ycf3



— 0.1 substitutions/site

Table IV.4. Individual phylogenetic analyses based on plastid-associated genes, including dinoflagellates.

Gene name	Analyses	Number of sites	Support values for different clades a											Other relationships b
			1	2	3	4	5	6	7	8	9	10	11	
<i>atpF</i>	TreePuzzle	68	-	-	-	-	-	-	-	-	-	-	-	-
<i>atpF</i>	PAUP	136	+	+	-	-	+	-	-	-	-	-	-	Odontella + green plastids
<i>ccsA</i>	TreePuzzle	109	-	-	-	-	-	-	-	-	-	-	-	-
<i>ccsA</i>	PAUP	218	+	-	-	-	+	-	-	-	-	-	-	-
<i>psaD</i>	TreePuzzle	138	-	-	68	-	-	-	-	-	50	-	-	-
<i>psaD</i>	PAUP	276	83	-	-	87	-	75	+	-	-	-	+	-
<i>psaI</i>	TreePuzzle	29	-	-	-	-	-	-	-	-	-	-	-	-
<i>psaI</i>	PAUP	58	-	-	-	-	-	-	-	-	-	-	-	Emiliania + Synechocystis
<i>psbI</i>	TreePuzzle	38	91	-	-	-	-	71	-	-	54	-	-	-
<i>psbI</i>	PAUP	76	83	-	-	-	-	63	-	-	-	-	-	-
<i>psbJ</i>	TreePuzzle	39	-	-	-	-	69	67	-	-	-	-	78	-
<i>psbJ</i>	PAUP	78	-	-	-	-	+	-	-	-	-	-	-	Synechocystis + Cyanophora
<i>psbZ</i>	TreePuzzle	62	91	-	-	-	85	-	-	-	-	-	-	-
<i>psbZ</i>	PAUP	124	100	-	-	-	+	-	+	-	-	-	65	-
<i>rpl14</i>	TreePuzzle	123	80	-	-	-	-	-	-	-	63	-	-	-
<i>rpl14</i>	PAUP	246	78	-	-	-	-	-	-	-	-	-	-	-
<i>rpl16</i>	TreePuzzle	132	82	-	-	74	89	89	57	-	64	-	-	-
<i>rpl16</i>	PAUP	264	+	-	+	-	+	+	-	-	-	-	57	-
<i>rpl19</i>	TreePuzzle	80	92	-	-	-	-	-	-	-	-	-	-	Emiliania + Mesostigma 74
<i>rpl19</i>	PAUP	160	+	-	-	-	-	-	-	-	-	-	-	-
<i>rpl20</i>	TreePuzzle	108	85	-	-	-	-	-	-	-	-	-	-	-
<i>rpl20</i>	PAUP	216	89	-	-	-	+	-	-	-	-	-	-	-
<i>rpl21</i>	TreePuzzle	101	-	-	-	-	-	-	-	-	-	-	-	Chaetosphaeridium + cyanobacteria
<i>rpl21</i>	PAUP	202	-	-	-	-	-	63	-	-	-	-	-	-
<i>rpl22</i>	TreePuzzle	103	-	-	-	-	-	-	-	-	-	-	-	Chaetosphaeridium + Cyanophora
<i>rpl22</i>	PAUP-no dino	206	-	-	-	-	-	-	-	-	-	-	-	-
<i>rps3</i>	TreePuzzle	145	57	-	-	53	-	-	-	-	-	-	-	-
<i>rps3</i>	PAUP-no dino	290	95	-	-	+	-	-	-	-	-	-	-	-
<i>rps4</i>	TreePuzzle	205	86	-	-	64	68	-	-	-	81	-	-	-
<i>rps4</i>	PAUP-no dino	410	50	-	-	-	-	-	-	-	+	-	-	-
<i>rps7</i>	TreePuzzle	155	81	-	-	59	85	54	-	-	-	-	-	-
<i>rps7</i>	PAUP-no dino	310	91	-	-	+	83	78	-	-	-	-	-	-
<i>rps8</i>	TreePuzzle	136	62	-	-	-	64	-	-	-	-	-	-	Porphyra + Cyanophora
<i>rps8</i>	PAUP-no dino	272	51	-	-	-	-	-	-	-	-	-	-	-
<i>rps11</i>	TreePuzzle	121	72	-	-	89	93	-	-	-	-	59	-	-
<i>rps11</i>	PAUP-no dino	242	63	-	-	63	77	+	-	-	-	+	-	-
<i>rps12</i>	TreePuzzle	122	85	-	-	-	-	-	-	-	59	-	-	Odontella + green plastids
<i>rps12</i>	PAUP-no dino	244	-	-	-	-	-	-	-	-	-	-	-	-
<i>rps14</i>	TreePuzzle	100	55	-	-	-	-	-	-	-	-	-	-	-
<i>rps14</i>	PAUP-no dino	200	-	-	-	+	-	-	-	-	-	-	-	-
<i>rps16</i>	TreePuzzle	71	79	-	-	68	64	-	-	-	-	-	-	Porphyra + green plastids
<i>rps16</i>	PAUP-no dino	142	-	-	-	+	-	-	-	-	-	-	-	-
<i>rps18</i>	TreePuzzle	62	72	-	-	-	-	-	-	-	61	-	-	-
<i>rps18</i>	PAUP-no dino	124	67	-	-	+	-	-	-	-	-	-	-	-
<i>rps19</i>	TreePuzzle	92	80	-	-	+	55	52	-	-	-	-	93	-
<i>rps19</i>	PAUP-no dino	184	57	-	-	-	-	+	+	61	+	-	-	-
<i>ycf4</i>	TreePuzzle	178	-	-	-	72	89	72	-	-	-	-	-	-
<i>ycf4</i>	PAUP-no dino	356	-	-	-	-	95	-	-	-	-	-	-	-

1- Cyanobacteria monophyly

2- Cyanophora + red plastids monophyly

3- Cyanophora + green plastids monophyly

4- Green + red plastids monophyly

5- Green plastids monophyly

6- Red plastids monophyly

7- Chl c containing plastids monophyly

8- Red algal plastids monophyly

9- Guillardia + Emiliania monophyly

10- Guillardia + Odontella monophyly

11- Emiliania + Odontella monophyly

TreePuzzle- Protein analyses based on JTT model of substitution using the program TreePuzzle

PAUP- Nucleotide maximum likelihood analysis based on GTR+I+G4 model of evolution using PAUP\*

a- Support values only when > 50. A dash (-) indicates that the clade is not present in the best tree

A plus (+) indicates that the clade is present in the best tree but the support value < 50.

b- Numbers correspond to support values for the relationships, when higher than 70

Figure IV.2. Maximum likelihood analyses based on individual plastid genes, under GTR+I+Γ4 using PAUP\*, without dinoflagellates. Numbers above branches are bootstrap support values, when > 60.

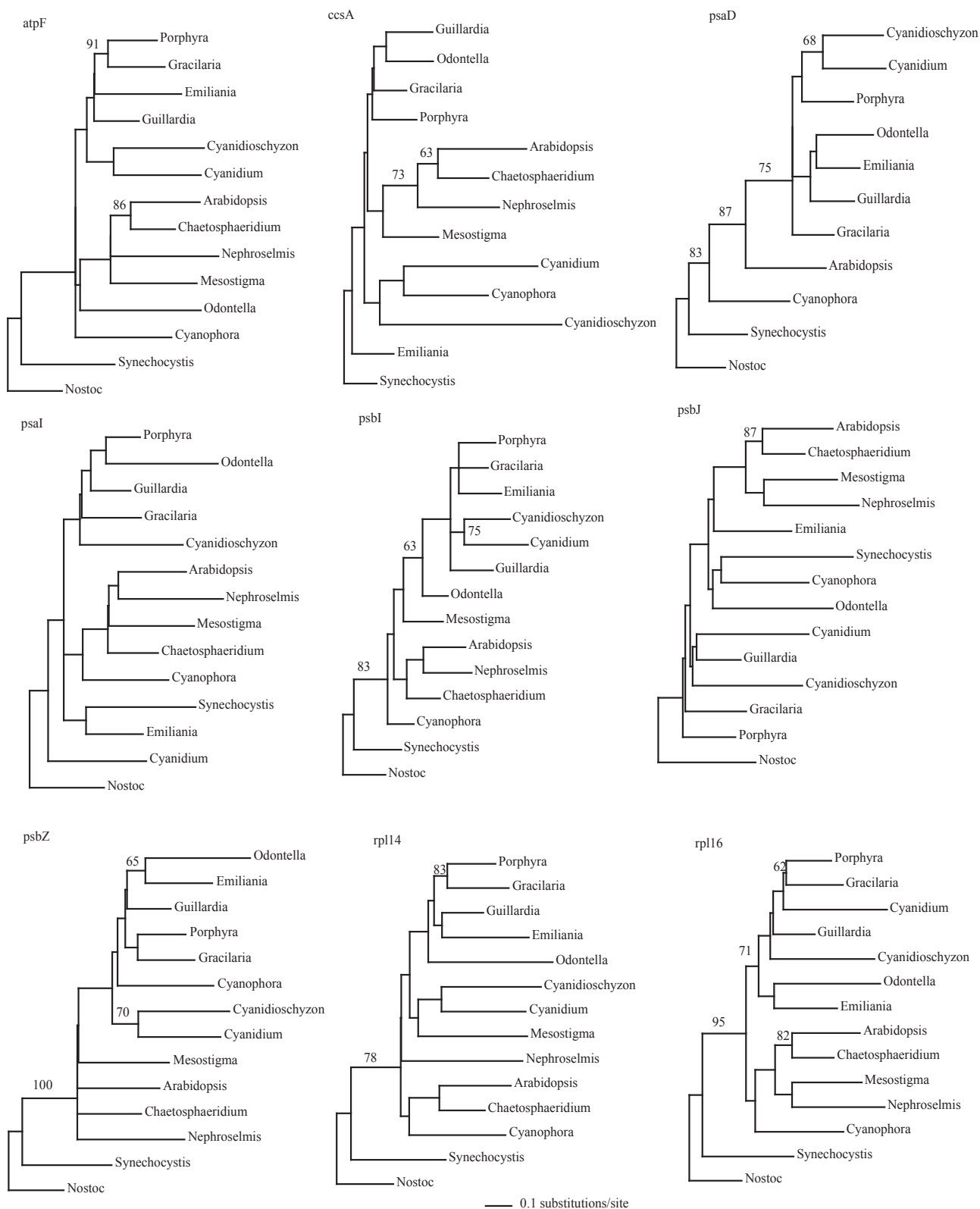


Figure IV.2 (con't). Maximum likelihood analyses based on individual plastid genes, under GTR+I+ $\Gamma$ 4 using PAUP\*, without dinoflagellates. Numbers above branches are bootstrap support values, when > 60.

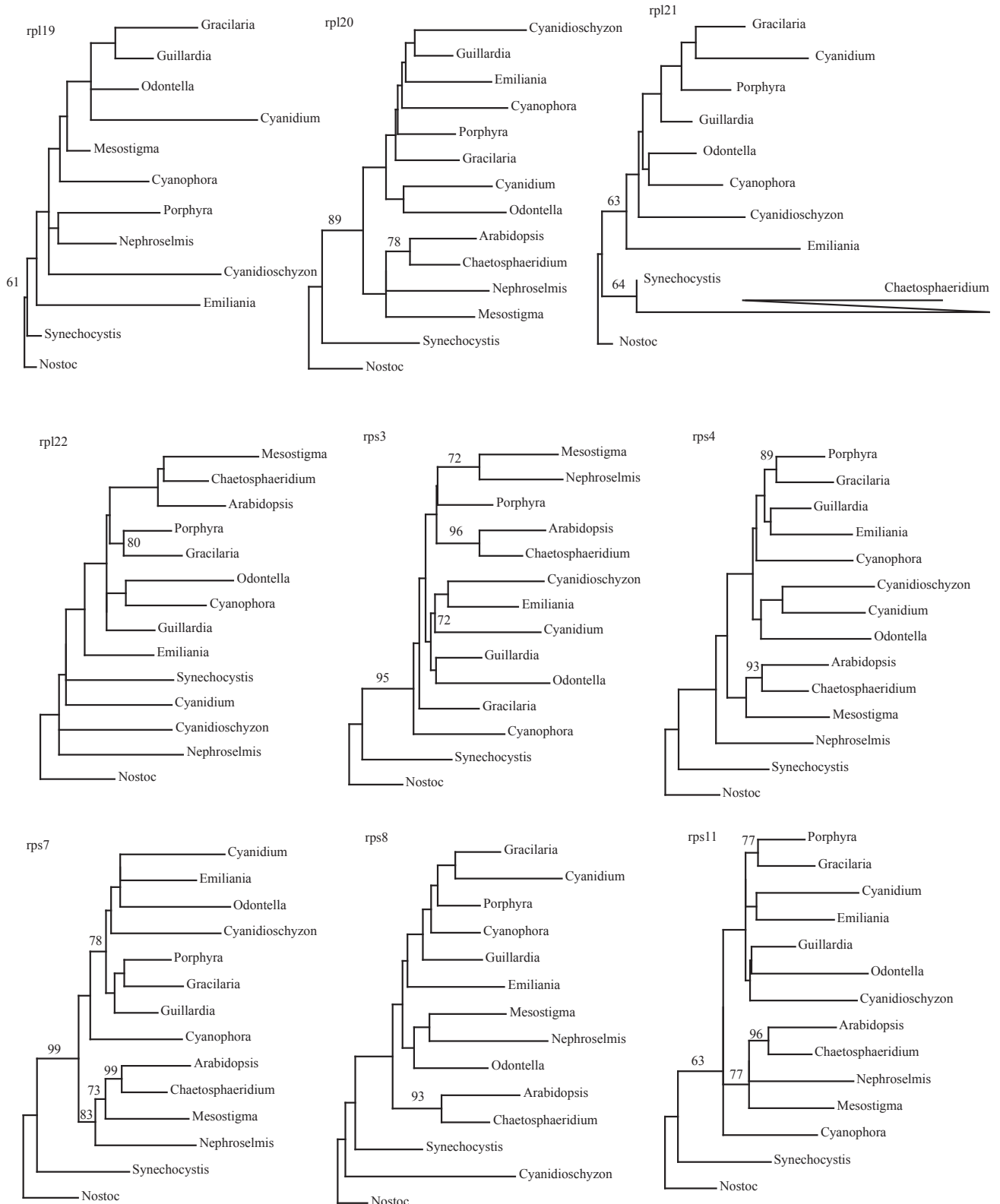
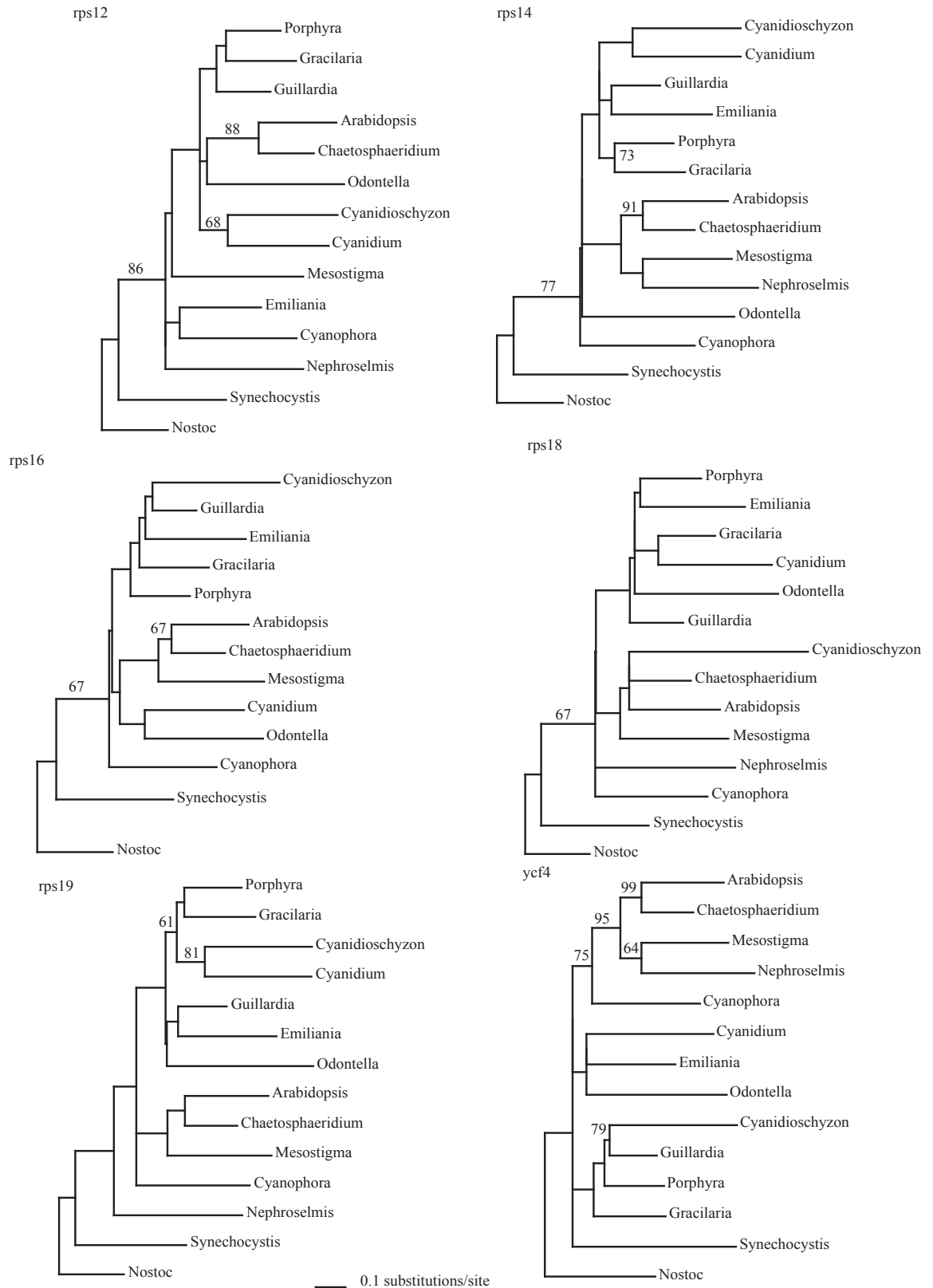




Figure IV.2 (con't). Maximum likelihood analyses based on individual plastid genes, under GTR+I+ $\Gamma$ 4 using PAUP\*, without dinoflagellates. Numbers above branches are bootstrap support values, when > 60.



Tables IV.3 and IV.4. The majority of these analyses find low support for most branches, although a few clades are well supported (Figures IV.1, IV.2). The cyanobacterial, green plastid, and red-lineage plastid clades are in general moderately or strongly supported (70-100% support values). When dinoflagellates are not included, a few gene trees show unexpected relationships in the most likely tree (Tables IV.3, IV.4). In all cases, these unexpected taxon bipartitions have low support (<70% support) and are found in only one type of analysis based on nucleotide or amino acid data.

Only 38 out of the 62 genes studied are available from peridinin-containing dinoflagellates. When dinoflagellates are included, the dinoflagellate plastid is associated in most single gene trees with the red-lineage plastid clade (Table IV.3). However, some analyses show a relationship of the dinoflagellate plastid with other taxa, in general with low support values (i.e. < 70%), and only recovered by one analytical method. In a few cases the support values for these unexpected relationships were moderate: *petA* (with *Nephroselmis*, 70 support value), *chlI* (with *Cyanophora*, 86), *psaC* (with *Cyanophora*, 71), *psaJ* (with *Synechocystis*, 70).

### **Concatenated gene analyses**

Several concatenated datasets (Table IV.5) were constructed and analyzed using amino acid and nucleotide based methods. Excluding dinoflagellates, two different datasets were considered. The most comprehensive dataset contains all the genes under study (62-gene dataset). I constructed a second dataset (24-gene dataset) in which the genes that show a higher rate of evolution as determined by Hagopian et

Table IV.5. Datasets based on plastid-associated genes used for the phylogenetic analyses.

Dataset	Number of nt <sup>a</sup>	Number of aac <sup>b</sup>	Taxa that failed the amino acid composition chi-square test <sup>c</sup>	Genes present in each dataset
Excluding dinoflagellates				
62 genes	23972	11986	<i>Arabidopsis</i> , <i>Cyanidioschyzon</i> , <i>Cyanidium</i> , <i>Gracilaria</i> , <i>Mesostigma</i> , <i>Nephroselmis</i> , <i>Nostoc</i> , <i>Odontella</i> , <i>Synechocystis</i>	All genes present in Tables IV.3 and IV.4.
24 genes	9002	4501	<i>Cyanidium</i>	<i>petA</i> , <i>petB</i> , <i>petD</i> , <i>petG</i> , <i>psaA/B/C/D/F/I/J/L</i> , <i>psbB/C/D/E/F/H/I/J/K/L/N/Z</i>
Including dinoflagellates				
38 genes	17458	8729	<i>Amphidinium</i> , <i>Chaetosphaeridium</i> , <i>Cyanidioschyzon</i> , <i>Cyanidium</i> , <i>Gracilaria</i> , <i>Nephroselmis</i> , <i>Nostoc</i> , <i>Synechocystis</i>	All genes present in Table IV.3. <i>AtpA/B/E/H/I</i> , <i>chlI</i> , <i>clpC</i> , <i>dnaK</i> , <i>petA</i> , <i>petB</i> , <i>petD</i> , <i>petG</i> , <i>psaA/B/C/F/J/L</i> , <i>psbB/C/D/E/F/H/K/L/N/T</i> , <i>rpl2</i> , <i>rpl3</i> , <i>rpl5</i> , <i>rpl23</i> , <i>rpl33</i> , <i>rps2</i> , <i>rps9</i> , <i>secA</i> , <i>tufA</i> , <i>ycf3</i>
15 genes	5694	2847	none	<i>petB</i> , <i>petD</i> , <i>petG</i> , <i>psaB</i> , <i>psaJ</i> , <i>psbB/C/D/E/F/H/K/L/N/T</i>

<sup>a</sup> Number of nucleotides (first and second codon positions) included in the phylogenetic analyses.

<sup>b</sup> Number of amino acids included in the phylogenetic analyses.

<sup>c</sup> In all datasets, at least five taxa failed the homogeneity chi-square test based on nucleotides.

Figure IV.3. Phylogenetic analyses based on plastid-associated genes, excluding the dinoflagellate. A, C. Trees based on the 62-gene dataset. B, D. Evolutionary trees based on the 24-gene dataset. A-B. Maximum likelihood analyses based on GTR+I+ $\Gamma$ 4 model of evolution. Bootstrap support values ( $>65$ ) from analyses excluding the third codon position and from analyses including only second codon position are shown above branches (on the left and right, respectively). Bootstrap values from LogDet distance analyses excluding third codon position are shown below the branches. C-D. Maximum likelihood protein analyses based on JTT model of amino acid substitution. Bootstrap support values ( $>65$ ) from the ProML analyses are shown above the branches, and quartet puzzling support values obtained using TreePuzzle are below the branches.

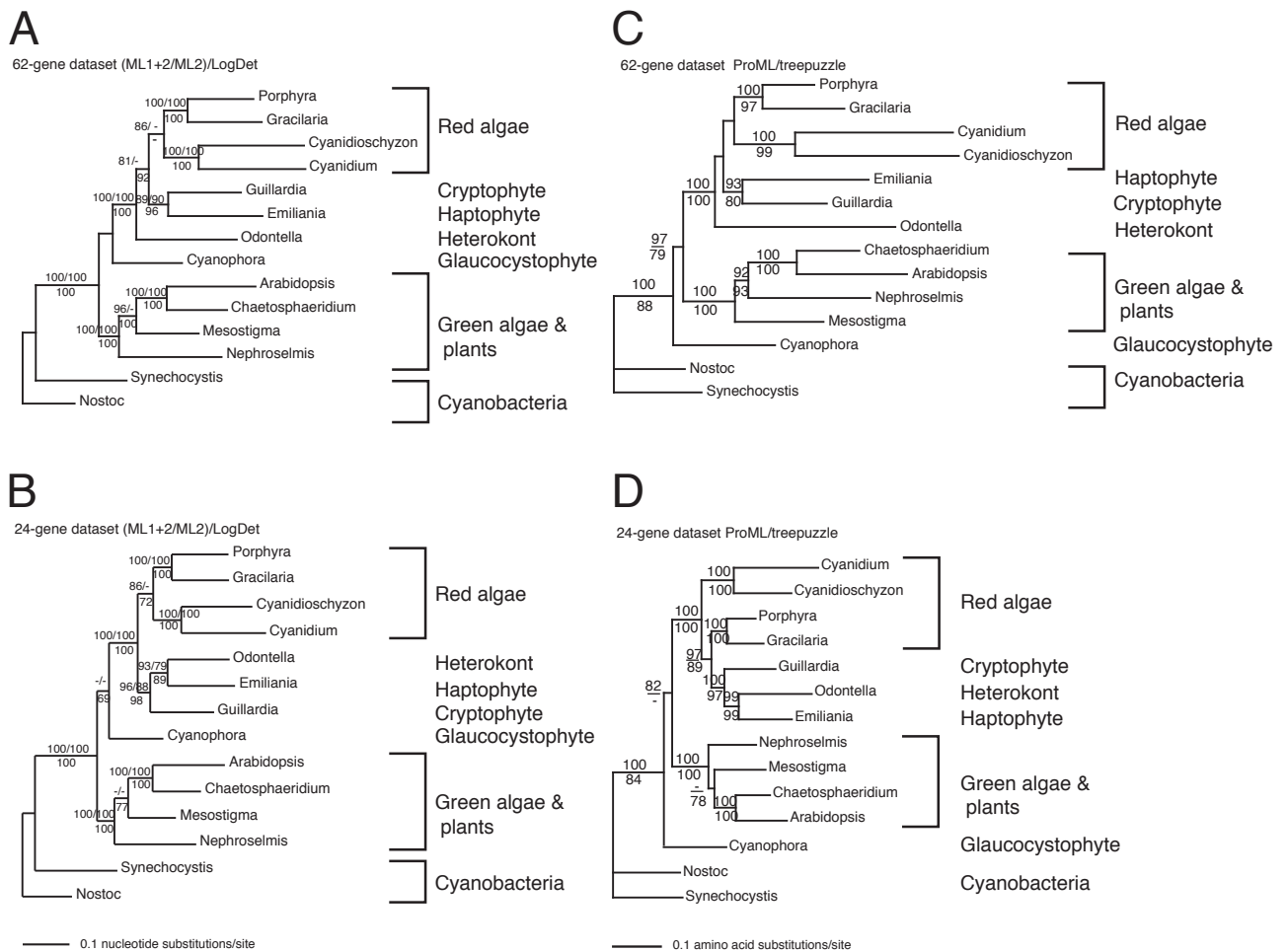


Figure IV.4. Phylogenetic analyses based on plastid-associated genes, including the peridinin-containing dinoflagellate (*Amphidinium* or *Alexandrium*). A, C. Trees based on the 38-gene dataset. B, D. Evolutionary trees based on the 15-gene dataset. A-B. Maximum likelihood analyses based on GTR+I+Γ4 model of evolution, excluding third codon position. Bootstrap support values are shown above branches when > 65. C-D. Maximum likelihood protein analyses based on JTT model of amino acid substitution. Bootstrap support values (> 65) from the ProML analyses are shown above the branches, and quartet puzzling support values obtained using TreePuzzle are below the branches.



al. (2004) were excluded and none one of the taxa (except *Cyanidium*) fail the compositional homogeneity chi square test of amino acid frequency distribution (Schmidt et al. 2002). I also performed analyses based on the 24-gene dataset excluding *Cyanidium* and the results (data not shown) were practically identical to those analyses including *Cyanidium*. Concatenated datasets including dinoflagellate sequences (Table IV.5) are limited by the data available (38 out of the 62 genes analyzed). One of the datasets (38-gene dataset) includes all of the available sequences. Another dataset (15-gene dataset) was constructed, in which the genes that show an extreme rate of sequence evolution leading to the dinoflagellate (Bachvaroff et al. 2006) were excluded, and all taxa passed the chi square test for amino acid frequency distribution.

All the analyses of these concatenated datasets found strongly supported clades, such as cyanobacteria, green plastids and red-lineage plastids (Figures IV.3, IV.4). In addition, some taxon bipartitions are present in all analyses with high support: *Arabidopsis* + *Chaetosphaeridium*, *Porphyra* + *Gracilaria*, *Cyanidium* + *Cyanidioschyzon*. Maximum Parsimony analyses (data not shown) agree in general with the results from the ML analyses. A red + green plastid clade is well supported by most datasets based on amino acid data. Nucleotide based analyses display a relationship of red-lineage plastids with *Cyanophora*, but in most cases the support is low.

Monophyly of red algal plastids is moderately supported by nucleotide ML analyses based on all datasets. On the other hand, red algal plastids are paraphyletic with respect to the chl c plastids in all amino acid based analyses under MP or ML,

whether or not the dinoflagellate is included. Given the incongruities between analyses based on nucleotides and amino acids (using the 62- and 24-gene datasets, in particular), LogDet distance analyses, and analyses including only the second codon position were performed. The nucleotide ML analyses based on the 62-gene and 24-gene datasets including only the second codon position, and the LogDet distance analysis based on the 62-gene dataset show the red algal plastids as paraphyletic (Figure IV.3). I also recoded the nucleotide data and amino acid data from the 62-gene dataset hoping to reduce the effect of compositional bias. The four-state nucleotide alignments were converted as RY-coding, while I converted all lysine residues into arginines (K=R) and all valine, leucine and methionine residues into isoleucines (I=V=L=M). The trees (data not shown) were not significantly different from the analyses of the original datasets.

Chl c containing plastids form a strongly supported clade using the 24-gene and 15-gene datasets, both in nucleotide and amino acid analyses (Figures IV.3, IV.4). In other analyses, where chl c plastids are not monophyletic (62- and 38-gene datasets), *Guillardia* and *Emiliana* form a highly supported monophyletic group, while *Odontella* is found sister to all red-lineage plastids with low to moderate support, or sister to a clade formed by *Porphyra*, *Gracilaria*, *Emiliana* and *Guillardia*. In the 38-gene nucleotide analysis, *Odontella* is sister to the dinoflagellate with low support. In amino acid MP analyses using the 62-gene dataset, chl c plastids are monophyletic and *Emiliana* is sister to *Odontella* with moderate support (data not shown). In the 24-gene analyses, where chl c plastids are monophyletic, *Odontella* is sister to *Emiliana* with strong support (93-99) and *Guillardia* is basal to this clade. In

the amino acid based 15-gene analysis, *Emiliana* forms a clade with the dinoflagellate with strong support (93), and this clade is sister to *Odontella* (92). The nucleotide-based 15-gene analysis shows *Emiliana* sister to *Odontella* with low bootstrap support (65), and this clade is sister to the dinoflagellate (61). In all analyses based on the 15-gene dataset, *Guillardia* is basal to all chl c plastids with low (61) to strong support (92).

### **Approximately Unbiased (AU) test**

The AU test was used to assess confidence in several phylogenetic hypotheses (Schimodaira 2002). In this test, a set of trees are constructed where individual nodes are constrained and the best tree compatible with this constraint is found. The site likelihoods of this set of trees are then compared using the AU test. No combination of glaucophyte, green, and red-lineage plastids is rejected by AU tests based on either nucleotide or amino acid analyses. Also, monophyly of red algal plastids and chl c containing plastid monophyly are not rejected by any analysis or dataset tested. A clade formed by *Porphyra*, *Gracilaria*, *Emiliana* and *Guillardia*, to the exclusion of *Odontella* and Cyanidiales is rejected by all datasets, except the 62-gene nucleotide dataset. All possible plastid associations within the chl c containing plastid clade were also tested. The only hypothesis that is widely rejected is a sister relationship of *Odontella* and *Guillardia*, which is rejected by the 62- and 24-gene datasets based on nucleotide or amino acid data and the 38- and 15-gene datasets based on nucleotides. The sisterhood of *Emiliana* + *Guillardia* is rejected by the 25-gene dataset based on amino acids. All other possible relationships are not rejected by any dataset.



## ***Discussion***

A heated debate exists regarding the relative importance of taxon sampling and site sampling in phylogenetic analyses, while limited resources force compromises between the number of genes vs. taxa included in the analyses. Large concatenated datasets reduce sampling error that can affect single or a few gene analyses, but run the risk of incorporating discordant data (Martin et al. 2005). Nonetheless, despite the desirability of large datasets and dense taxon sampling, real analyses are limited to the available data. Here, I analyzed large datasets derived from complete plastid genomes and nuclear-encoded plastid-targeted genes to study plastid evolution in several eukaryotic lineages (Table IV.1). In an attempt to distinguish phylogenetic signal from noise, a broad range of analyses were performed, with bootstrapping and AU tests to assess confidence in the results.

I observed that combining a high number of plastid-associated genes with dissimilar evolutionary rates affects the inference of evolutionary relationships, as does inclusion of taxa with biased amino acid and nucleotide frequency distribution. Analyses based on more consistent datasets (see below) suggest that secondary plastids from chl c containing algae were acquired from red algae after the divergence of Cyanidiales from the other red algae, and that chl c containing plastids form a monophyletic group. Novel findings include the position of the cryptophyte plastid basal to the chl c plastid clade, and the peridinin-containing dinoflagellate plastid sister to the haptophyte plastid. Overall, to the best of my knowledge, these results represent the largest presently available dataset able to examine the relationships among all chl c containing plastids.

## Individual gene analyses

In general, single gene phylogenies are not strongly supported, presumably because of the small number of characters included in each analysis. In most cases the best tree showed cyanobacteria, green plastids, and red-lineage plastids to be monophyletic with moderate to strong support, while relationships within those clades were typically unresolved. In general, I did not find significant incongruence between single-gene phylogenies.

Morphological and ultrastructural data lead most investigators to infer that the peridinin-containing dinoflagellate plastid is derived from the red algae (Gibbs 1978; Cavalier-Smith 1999; Fast et al. 2001). However, in single gene analyses the dinoflagellate plastid often grouped with green plastids, the glaucophyte plastid, or cyanobacteria. This could be explained by lateral gene transfer from a green alga, glaucophyte, or cyanobacteria to the dinoflagellate (Hackett et al. 2004), or it could be an artifact of the data or the analytical method. To distinguish between these possibilities, I further analyzed the genes that found unexpected relationships. First, I observed that most of these relationships (11 out of 18) are found in only one type of analysis (i.e. nucleotide but not amino acid analyses of the same gene, or *vice versa*). In addition, in five cases the taxa involved in the unexpected relationship violated the underlying assumption of symmetric substitution. For example, *rpl5*, *rps9*, and *atpH* showed the dinoflagellate plastid sister to the green or glaucophyte plastid with moderate support, and in those cases, the dinoflagellate sequence failed the amino acid frequency distribution chi square test. AU tests were performed on these genes,

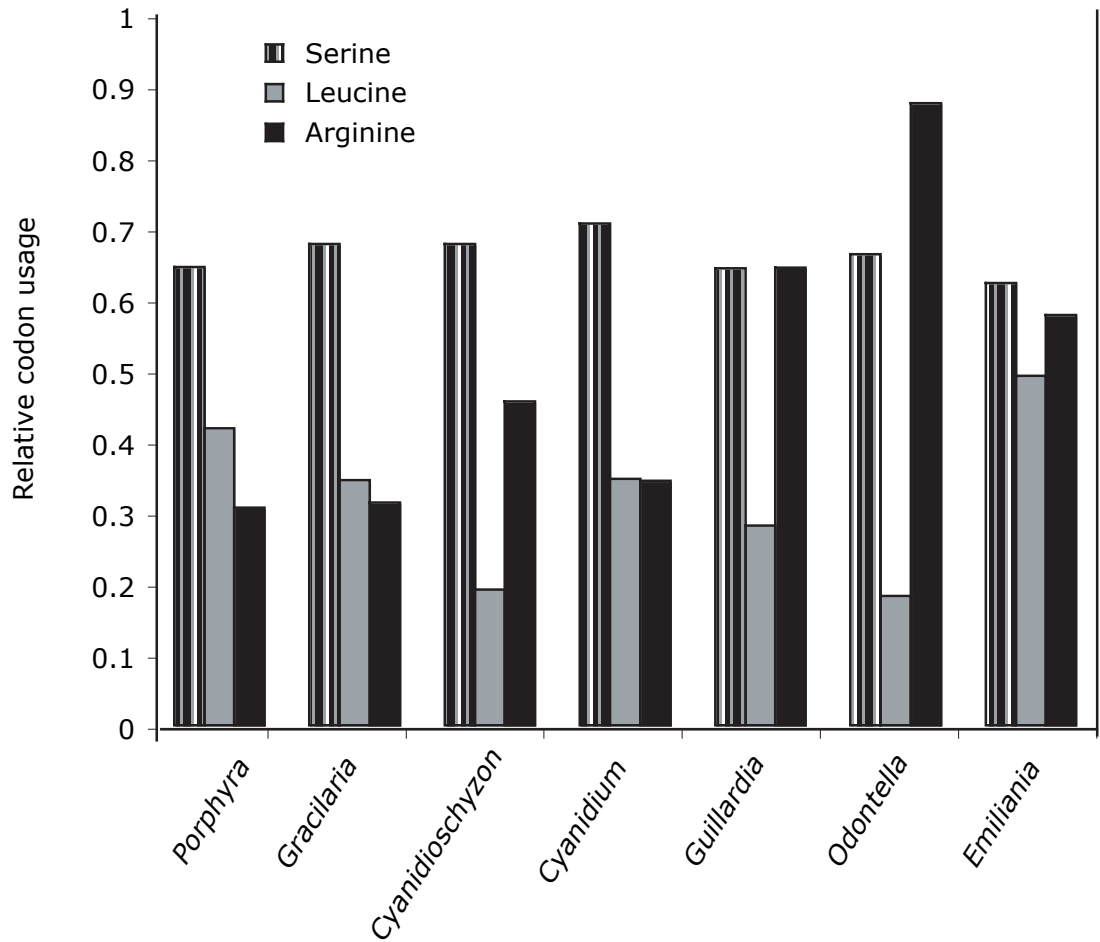
and none of them rejected a clade associating the dinoflagellate plastid with red-lineage plastids. Similarly, the *clpC* tree showed a dinoflagellate + *Synechocystis* clade, with *Synechocystis* falling among the red plastids, and the *Synechocystis clpC* sequence failed the chi square test. That is also true with *Chaetosphaeridium* in the *tufA* analysis, with *Chaetosphaeridium* lying among the red plastids, sister to the dinoflagellate plastid, and failed the homogeneity test. Therefore, the apparent close association of some dinoflagellate plastid genes with green or glaucophyte genes can be explained by analytical artifacts, and may not be the result of horizontal gene transfer.

### **Reliability of different datasets and analytical methods**

A number of concatenated datasets were constructed and analyzed in this study, using both amino acid and nucleotide data. Substantial contradictory relationships were observed to be a function of analytical methods and type of data. For example, the 24-gene dataset showed a moderately supported red algal plastid clade in the nucleotide analysis, while protein analysis found a strongly supported plastid clade formed by two red algal plastids from *Porphyra*, *Gracilaria*, and the chl *c* plastids (Figure IV.3). This was also true when *Cyanidium* was excluded from the analyses (data not shown).

To understand organismal phylogeny, it is essential to understand the basis for such apparent conflict from different data types and analytical methods. One explanation for differences in the trees found by amino acid vs. nucleotide data could

Figure IV.5. Relative codon usage for serine (TCN/TCN+AGY), leucine (CTN/CTN+TTR) and arginine (CGN/CGN+AGR) observed in seven taxa based on the 24-gene dataset. Red algal plastids (from *Porphyra*, *Gracilaria*, *Cyanidioschyzon* and *Cyanidium*) share a common codon usage for arginine different from the one used by chl c containing plastids (from *Guillardia*, *Odontella*, and *Emiliana*).



be compositional bias. Synonymous substitutions in nucleotide data lead to higher evolutionary rates, and the affected positions can saturate rapidly with a consequent risk of showing a misleading phylogenetic signal. A common approach to this problem, particularly if the sequences also display some compositional bias, is to exclude the third codon position from nucleotide analyses, but synonymous substitutions can also occur in first codon positions with the same frequency, and these could have an important effect on the phylogenetic analysis (Inagaki et al. 2004a). Therefore, I performed ML analysis including only the second codon position, which is less susceptible to multiple substitutions than the other two, but reduces the total information available by roughly two thirds. In this analysis, the tree did not show a red algal plastid clade, in agreement with the amino acid based tree.

Another way to compensate for compositional bias is to use LogDet distance analyses. LogDet/Paralinear distance analyses are known to be robust to heterogeneous base composition (Lake 1994). However, the 24-gene LogDet tree showed a monophyletic red algal clade when first and second codon positions were included (Figure IV.3). It seems that LogDet could not deal correctly with the compositional heterogeneity under these particular conditions. The reasons why LogDet sometimes fails to recover the expected tree topology from compositionally heterogeneous alignments are still unclear (Jermiin et al. 2004). I also examined the codon usage of the amino acids that are susceptible to synonymous substitutions in the first codon position: leucine, serine and arginine. Inagaki et al (2004a) showed that codon usage heterogeneity affected the inference of evolution using *psbA* gene because of an unusual pattern in the first codon position shared by two unrelated

lineages. I performed similar tests to characterize the bias in the 24-gene dataset. Analyses of compositional bias showed that *Porphyra*, *Gracilaria*, and the Cyanidiales share a common codon usage pattern for arginine (Figure IV.5). This bias could explain the strongly supported monophyly of the red algal plastids found by the nucleotide based analysis and not shown in the amino acid based analysis.

Phylogenetic methods assume a symmetric substitution model, and are not particularly robust to an asymmetric (biased) substitution processes (Lockhart et al. 1999; Phillips et al. 2004; Martin et al. 2005). The program TreePuzzle performs a chi-square test on the data and compares the amino acid or nucleotide composition of each sequence to the frequency distribution assumed in the model (Schmidt et al. 2002). Analyses of compositional bias of nucleotide data (including or excluding the third codon position) showed that at least five taxa from each dataset rejected the homogeneity test. Compositional bias can also affect protein data and is known to have a strong influence on phylogenetic inference (Phillips et al. 2004). Thus, a dataset where this assumption is not violated would be preferable to one with known compositional bias. In most of the datasets investigated, the amino acid composition of the proteins from *Nostoc*, *Synechocystis*, *Gracilaria*, *Cyanidioschyzon*, *Cyanidium*, and *Nephroselmis* differed at  $p = 0.05$  from the expected frequency distribution (Table IV.5). In two of the datasets presented here (24-gene and 15-gene dataset), none of the sequences violate this assumption (except *Cyanidium* in the 24-gene dataset; however, analyses excluding this taxon found similar trees); and thus, the trees found by the analyses based on these datasets might be expected to be more reliable than the others.

## **Relationships among glaucophyte, red, and green plastids and their host cells**

Relationships among the three primary plastids and their host cells have been widely investigated, with multiple nuclear or mitochondrial gene analyses showing contradictory results. One hypothesis states that multiple independent primary endosymbiotic events took place in evolution, and glaucophytes, red, and green algae are not closely related to each other (Bhattacharya et al. 1995; Nozaki et al. 2003b; Stiller et al. 2003). Alternative associations of these three lineages with other eukaryotic groups have been recovered but in all cases they were poorly supported (Bhattacharya et al. 1995; Van de Peer and De Wachter 1997; Baldauf et al. 2000). Another hypothesis states that only one primary endosymbiosis gave rise to the Plantae, comprising glaucophytes, red, and green algae (Keeling 2004). In support of this hypothesis, a number of studies based on multiple nuclear or mitochondrial genes showed that green and red algae form a monophyletic clade called Plantae or Archaeplastida (Delwiche et al. 1995; Burger et al. 1999; Moreira et al. 2000; Philippe et al. 2004; Sanchez-Puerta et al. 2004), with the placement of the glaucophytes less certain. A recent study based on 143 nuclear genes from 34 eukaryotes recovered a strongly supported Plantae clade (Rodriguez-Ezpeleta et al. 2005), but the taxon sampling was poor and did not include most of the conflicting lineages.

If we assume a single primary endosymbiotic event prior to the divergence of glaucophytes, red, and green algae, analyses of plastid genes would not only describe

the evolution of primary plastids, but also host cell phylogenetic relationships. In the present study, most amino acid analyses, including the 24 and 15-gene datasets, found trees with high support for a clade formed by green and red-lineage plastids, with the glaucophyte plastid at the base. Therefore, the results presented here are consistent with a single endosymbiotic event before the divergence of these three algal lineages where glaucophytes branched before the green and red algae evolved (Figure IV.6). However, these data test only plastid relationships, and analyses including a wider range of cyanobacteria might show that primary plastids from these three lineages were acquired in independent primary endosymbioses from different cyanobacteria. And even if the three primary plastid lineages are all monophyletic, that doesn't guarantee that they are the result of a single endosymbiotic event.

### **Other observations: green plastid phylogeny**

The placement of the green alga *Mesostigma* remains unknown. Ultrastructural characteristics place it with the charophytes; however, molecular data supported conflicting phylogenetic relationships. Previous phylogenetic analyses showed support for different affiliations: either *Mesostigma* emerged before the divergence of chlorophytes and streptophytes (Lemieux et al. 2000; Hagopian et al. 2004), or it was sister to the charophyte algae and embryophytes (Karol et al. 2001; Martin et al. 2002). The latter hypothesis is consistent with the results presented here. In this study, the position of *Mesostigma* varied depending on the analyses and the dataset used (Figures IV.3, IV.4), but most analyses displayed strong support for a clade formed by *Arabidopsis*, *Chaetosphaeridium* and *Mesostigma*, with



*Nephroselmis* basal to the green plastid clade. Also, the 62- and 15-gene datasets based on nucleotides rejected a basal placement of *Mesostigma* within the green plastid clade.

### **Red algal plastids are monophyletic or paraphyletic?**

It is an open question whether the chl c containing algae acquired their plastids before or after the divergence of the Cyanidiales from the rest of the rhodophytes (Yoon et al. 2002b; Nozaki et al. 2004; Yoon et al. 2005). If the secondary endosymbiotic event(s) between chl c containing algae and rhodophytes occurred before the diversification of extant groups of red algae, then, the red algal plastids known today should form a monophyletic clade. Previous analyses based on a concatenated plastid-encoded or nuclear-encoded plastid-targeted genes found contradictory results (Yoon et al. 2002b; Nozaki et al. 2004; Yoon et al. 2005). In the present study, amino acid based analyses did not find red algal plastids to be monophyletic and showed moderate to strong support for a clade formed by chl c plastids and *Porphyra* + *Gracilaria*. In the nucleotide based analyses I detected artificial results due to compositional bias (see above). Overall, these results suggest that the chl c containing algae acquired their plastids after the divergence of Cyanidiales from other red algae and before the divergence of members of the class Bangiophyceae (*Porphyra*) and those of the class Florideophyceae (*Gracilaria*, Figure IV.6). Additional analyses including wider taxon sampling of red algae would describe more accurately the source of the chl c containing plastids.

### **Are chl c containing plastids monophyletic?**

The monophyly of the chl c containing plastids has been evaluated using both single and multi-gene phylogenies, and also by correlation analyses of complete plastid genomes with results that were often times contradictory. Analyses of a few concatenated plastid genes showed support for monophyly of the chl c plastids (Yoon et al. 2002b; Bachvaroff et al. 2005; Yoon et al. 2005). In contrast, analyses based on much larger datasets of plastid genes, but including two chl c plastid lineages, did not find a chl c plastid clade (Martin et al. 2002; Hagopian et al. 2004). Nuclear-encoded, plastid-targeted gene analyses argued in favor of a single origin of chl c containing plastids (Fast et al. 2001; Ishida and Green 2002; Harper and Keeling 2003; Patron et al. 2004). In the present study, monophyly of chl c containing plastids was highly supported by the most conservative datasets, 24 and 15-gene analyses using both nucleotide and amino acid data. In the analyses where these plastids were not found to be monophyletic, the support for other relationships was low and none of the AU tests based on these data rejected their monophyly. Therefore, the results shown here suggest monophyly of the chl c containing plastids derived from the red algae (Figure IV.6A). However, this information does not directly address the number of endosymbiotic events that took place in the evolution of these lineages.

### **Are chl c containing host cells monophyletic?**

A monophyletic chl c plastid clade (Figure IV.6A) is congruent with at least two different host cell evolutionary hypotheses (Figure IV.6B, C). One possibility is that chl c containing host cells are also monophyletic (“chromalveolate hypothesis”,

Figure IV.6. Diagram of evolutionary hypotheses of photosynthetic eukaryotes and their plastids. A. Plastid evolution hypothesis as suggested by the data analyzed in the present study. B-C. Two alternative hypotheses of the evolution of photosynthetic eukaryotes. Both alternative hypotheses of host cell evolution are congruent with the plastid evolution shown in A. Other hypotheses can also be postulated. B. Chromalveolate hypothesis; chlorophyll c containing algae are monophyletic and they acquired the plastid from a red alga after the evolution of the Cyanidiales, in a single endosymbiotic event prior to their divergence. Crossed circles indicate that those lineages lost their ability to photosynthesize and probably their plastids. C. Serial hypothesis; chl c containing lineages are not closely related and they acquired their plastids in independent endosymbiotic events. Shown here is a particular case of this hypothesis where four endosymbioses took place in the evolution of the four lineages of chl c algae with a single acquisition from the red algae after the divergence of the Cyanidiales.

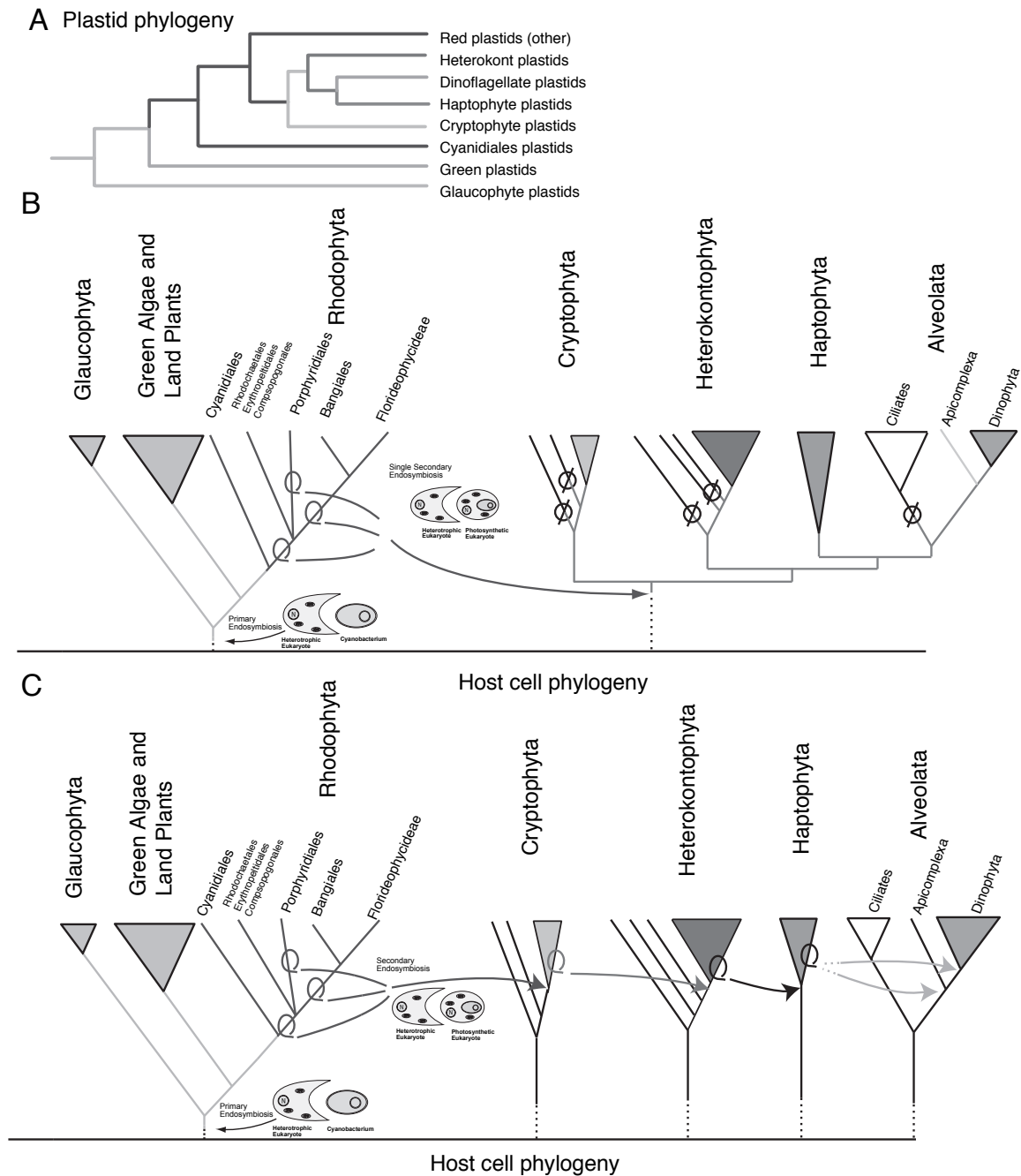


Figure IV.6B), and only a single secondary endosymbiosis with a red alga had occurred prior to the divergence of the four chl c containing lineages (Cavalier-Smith 1986; Cavalier-Smith 1999). Under this assumption, several heterotrophic lineages that diverged early in the evolution of cryptophytes, heterokonts, and alveolates lost their plastids and are considered secondarily heterotrophic (e.g., ciliates). Many scientists remain skeptical about the chromalveolate hypothesis, in part because molecular data do not support the monophyly of the host cells (Medlin et al. 1997; Sanchez-Puerta et al. 2004; Harper et al. 2005), and also because analyses of the nuclear genome of the heterotrophic oomycete *Phytophthora infestans* and the ciliate *Tetrahymena thermophila* did not find any plastid-derived genes (but see (Andersson and Roger 2002) as predicted by this hypothesis.

Alternative hypotheses can also be postulated, describing independent endosymbiotic events among chl c containing organisms (Cavalier-Smith et al. 1994; Delwiche 1999; Bachvaroff et al. 2005). For example, the “serial hypothesis” states that the host cells are not closely related, and that the plastids that derived from the red algae were passed on among the chl c containing lineages in 2-4 separate endosymbiotic events. This hypothesis is also consistent with the plastid evolution described in Figure IV.6A, where chl c containing plastids share a common ancestor from the red algae. A specific case of this hypothesis is shown in Figure IV.6C (several variants of this hypothesis could also be proposed), where a cryptophyte engulfed a red alga and kept it as a permanent endosymbiont. In this hypothesis, a heterokont subsequently acquired the plastid from a cryptophyte in a tertiary endosymbiotic event, and even later, a haptophyte engulfed a heterokont. The

dinoflagellate plastid is also assumed to have been derived from a haptophyte. This hypothesis is quite complex because it invokes several independent endosymbiotic events, but it is not necessarily less parsimonious than a single endosymbiotic event given current data on host and plastid relationships. A permanent establishment of the endosymbiont in the host cell environment requires a number of processes to occur, such as developing a mechanism for importing proteins into the plastid, and acquisition of signal and transit peptides by the nuclear-encoded plastid-targeted genes (Cavalier-Smith 1999). Because the mechanisms by which such adaptations occur are poorly understood, it is impossible to model expectations for their probability at this time.

### **Relationships among chl c containing plastids**

Relationships within the chl c containing plastid clade are still uncertain, especially when the rapidly evolving peridinin-containing dinoflagellate sequences are included. Previous analyses based on 5 or 9 plastid genes showed the cryptophyte plastid basal to the chl c plastid clade and a sisterhood of *Emiliana* + *Odontella* (Yoon et al. 2002b; Bachvaroff et al. 2005). In contrast, other analyses based on plastid-encoded or nuclear-encoded plastid-targeted genes showed the haptophyte plastid as sister to other chl c plastids (Daugbjerg and Andersen 1997; Harper and Keeling 2003; Patron et al. 2004). Here, when the dinoflagellate is not included, the 62-gene dataset supports the sisterhood of *Emiliana* + *Odontella*, and the 24-gene analysis, which may be more reliable (see above), supports the relationship *Emiliana*

+ *Odontella*, with *Guillardia* sister to that clade, and rejects the clade formed by *Emiliana* + *Guillardia*.

Phylogenetic relationships of the peridinin-containing dinoflagellate plastid are difficult to resolve due to the high rate of evolution observed in many plastid-associated genes in this group (Zhang et al. 1999; Zhang et al. 2000; Bachvaroff et al. 2006). Previous multigene analyses based on plastid-encoded genes showed *Guillardia* at the base of the chl c clade, but the relationships among heterokont, haptophyte and dinoflagellate plastids were not well supported (Durnford et al. 1999; Ishida and Green 2002; Bachvaroff et al. 2005; Yoon et al. 2005). One analysis showed strong support for haptophyte + dinoflagellate plastids (Yoon et al. 2002a), but this result was based on the gene *psbA* which was later shown to have a misleading phylogenetic signal when using current analytical methods (Inagaki et al. 2004a; Bachvaroff et al. 2005). Phylogenetic analyses based on nuclear-encoded plastid-targeted genes, where the dinoflagellate sequences have a rate of evolution similar to the other taxa, found the haptophyte plastid at the base of the chl c containing plastid clade with moderate support (Harper and Keeling 2003; Patron et al. 2004). In the present study, the 15-gene protein analysis showed a strong association of the dinoflagellate with *Emiliana* (93), this clade sister to *Odontella* with high support (92) and *Guillardia* basal to the clade (Figures IV.4, IV.6A).

## ***Conclusions***

Phylogenetic inference based on sequence data intends to recover the genuine evolutionary history given the assumptions implicit in the alignment and in the

analytical method. However, signals other than the historical one can mislead analytical methods particularly when there are violations of the assumptions of the method. Under most common analytic methods in use today, it is difficult, for example, to accommodate highly compositionally biased datasets. (Martin et al. 2005). Covarion (for proteins, covariotide for nucleotides) evolutionary patterns, as well as heterogeneous evolution (asymmetrical substitutions leading to compositional heterogeneity among lineages) have been shown for other plastid gene datasets (Lockhart et al. 1999; Ane et al. 2005; Martin et al. 2005). Models that are flexible enough to accommodate deviations from traditional assumptions are often parameter-rich, and therefore computationally difficult and often lacking in power. It is therefore desirable to use the simplest method that is consistent with the characteristics of the data (Edwards 1972). In this paper, I tested the data for compositional bias, applied different methods to compensate for it (LogDet distance estimates, recodification of the data, use of amino acids), and reduced the number of useable genes from 62 to 24 to analyze a more homogeneous dataset. These analyses indicate that chlorophyll c containing plastids are monophyletic, and were acquired from the red algae after the divergence of the Cyanidiales; that *Guillardia* is basal to this clade; and that the haptophyte plastid seems to be sister to the dinoflagellate plastid. These data also indicate that any conclusions using current data and phylogenetic analyses should be considered provisional pending substantially more thorough taxon sampling, assembly of large, multi-gene datasets, and application of analytical methods that can accommodate heterogeneity in the data.

## **Chapter V – The Heterotrophic Dinoflagellate**

### ***Crypthecodinium cohnii* Descends from a Plastid-bearing Ancestor, Suggesting an Earlier Acquisition of Plastids**

#### ***Abstract***

Dinoflagellates are a diverse group of protists, comprising photosynthetic and heterotrophic free-living species, as well as parasitic ones. About half of them are photosynthetic with peridinin-containing plastids being most common. The peridinin-type plastid has been lost and replaced many times by other plastid types. Among dinoflagellates, photosynthetic species form a derived monophyletic clade while basal lineages are heterotrophic. It has been suggested that plastid acquisition occurred in the common ancestor of all photosynthetic dinoflagellates after the divergence of basal heterotrophic ones. An alternative hypothesis proposes an earlier plastid acquisition before the divergence of dinoflagellates and three other algal groups, namely cryptophytes, haptophytes, and heterokonts. Both hypotheses are consistent with the data available today. Studies of heterotrophic species from these lineages may increase our understanding of plastid evolution. I analyzed an EST project on the early-divergent heterotrophic dinoflagellate *Crypthecodinium cohnii* looking for evidence of past endosymbiosis. A significant number of genes of cyanobacterial or algal origin were identified using the BLAST tool from NCBI. Proteins known to be involved in plastid functions were used to directly search the *C. cohnii* database. Phylogenetic analyses suggest that several proteins could have been acquired from a



photosynthetic endosymbiont, arguing for an earlier plastid acquisition event. In addition, a putative N-terminal targeting signal was detected, indicating that *C. cohnii* may contain a reduced plastid and that some of these proteins are imported into this organelle. A number of metabolic pathways, such as heme and isoprenoid biosynthesis, and iron-sulfur cluster assembly, seem to take place in the plastid. This represents the first extensive genomic analysis of a heterotrophic dinoflagellate.

## ***Introduction***

Dinoflagellates are biflagellate, unicellular eukaryotes that live in marine and freshwater environments. Although about half of all dinoflagellates are heterotrophic, previous genome scale analyses of dinoflagellates have been limited to photosynthetic members. Apicomplexans (e.g., *Plasmodium*, the causative agent of malaria) and ciliates are dinoflagellates' closest known relatives and they are collectively called Alveolates, based on the cortical alveoli shared by these three lineages (Cavalier-Smith 1991). Alveolates themselves may be sibling taxon to the Chromists *sensu* Cavalier-Smith (1991), constituting a major eukaryotic lineage termed Chromalveolates (Cavalier-Smith 1999). Dinoflagellates form an ecologically diverse group, including photosynthetic, heterotrophic, and mixotrophic free-living species, as well as parasitic ones (Schnepf and Elbrächter 1992). Different dinoflagellate species host a variety of plastid types, and are photosynthetic. The most common plastid contains peridinin as the main accessory pigment. Plastids from peridinin-containing dinoflagellates, heterokonts, haptophytes, and cryptophytes (collectively called “chl c plastids”) are derived from the red algal lineage, contain chlorophyll *c*, and form a monophyletic clade (Yoon et al. 2002b; Bachvaroff et al. 2005). Other

lineages of dinoflagellates replaced the peridinin-type plastid with another one acquired from a green alga, a cryptophyte, a haptophyte, or a diatom (Cavalier-Smith 1999; Delwiche 1999; Bhattacharya et al. 2003). Molecular studies have shown that photosynthetic dinoflagellates descend from a common ancestor while several heterotrophic dinoflagellate lineages are basal to this clade (Gunderson et al. 1999; Saldarriaga et al. 2001; Leander and Keeling 2004). In addition, presumably heterotrophic lineages sister to the photosynthetic dinoflagellate clade have been identified through environmental sampling (Lopez-Garcia et al. 2001). The monophyletic origin of all photosynthetic dinoflagellates led to the idea that plastid acquisition in this group occurred in the common ancestor of all photosynthetic dinoflagellate species after the divergence of the basal heterotrophic lineages (Saunders et al. 1997; Saldarriaga et al. 2001). Under this hypothesis, early-divergent heterotrophic dinoflagellates would never have contained a plastid. Apicomplexans, sister taxa to the dinoflagellates, possess a non-photosynthetic plastid (apicoplast) with a remnant plastid genome (Foth and McFadden 2003). It is still controversial whether the source of this organelle is derived from a green or red algal endosymbiont (Funes et al. 2004). Strong evidence for a red ancestry of the apicoplast comes from studies of the gene GAPDH (glyceraldehyde 3-phosphate dehydrogenase) (Fast et al. 2001).

Acquisition of organelles is a complex process that involves the integration of both endosymbiont and host genomes. The endosymbiont could be a cyanobacterium, or a photosynthetic eukaryote engulfed by a heterotrophic eukaryote in a primary or secondary endosymbiotic event, respectively (Delwiche 1999; Bhattacharya et al.

2003). Upon endosymbiosis, a number of genes from the endosymbiont are transferred to the nuclear genome of the host cell (Martin et al. 1998). In general, protein products of the genes transferred to the nucleus are targeted back to the plastid aided by an N-terminal targeting signal (Kroth 2002). Considering the complexity of this process, a minimum number of endosymbiotic events giving rise to all plastids has been postulated; however, many of these lineages, including dinoflagellates, have nonphotosynthetic members (Cavalier-Smith 1981; Cavalier-Smith 2002). A single plastid acquisition for all chl c containing algae would indicate that nonphotosynthetic lineages had once contained a plastid. To the best of my knowledge, no evidence of past endosymbiosis has been found in the nucleus of ciliates, or in basal heterotrophic dinoflagellates. Genome analysis of those heterotrophic organisms would be beneficial to elucidate the evolution of the group and the estimated time of plastid acquisition.

*Cryptothecodinium cohnii* is an early divergent, non-photosynthetic marine dinoflagellate. As an osmotroph it can be easily cultivated by adding glucose and acetate as primary carbon sources (Javornicky 1962; Tuttle and Loeblich 1975). This species is important for its production of docosaheptaenoic acid (DHA), which is being used in commercial processes (Sijtsma and Swaaf 2004). Evolutionary relationships of *C. cohnii* are currently under study. Phylogenetic analyses show *C. cohnii* sister to the photosynthetic dinoflagellate clade (Saunders et al. 1997; Litaker et al. 1999; Saldarriaga et al. 2003; Leander and Keeling 2004; Zhang et al. 2005), or embedded in a clade with photosynthetic dinoflagellates (Saldarriaga et al. 2001; Saldarriaga et al. 2003; Murray et al. 2005), but in most cases the phylogenetic

position is essentially unresolved. Some authors believe, based on morphological data, that *C. cohnii* is a member of the Gonyaucales (Fensome et al. 1999; Saldarriaga et al. 2001; Murray et al. 2005), and that it is secondarily heterotrophic (Saldarriaga et al. 2001). In contrast, recent phylogenetic analyses based on the mitochondrial gene cytochrome b (*cob*) show *C. cohnii* outside the clade of plastid-containing dinoflagellates with moderate to strong support (Zhang et al. 2005).

In the present study, I analyzed an EST (expressed sequence tag) library from *C. cohnii* to look for evidence of past endosymbiosis. The first goal was to search for putative plastid-associated genes in the nuclear genome of *C. cohnii*. Plastid-associated genes are defined here as genes that are plastid-encoded, plastid-targeted nuclear-encoded, or nuclear-encoded genes that originated in a photosynthetic endosymbiont that may or may not be plastid-targeted. I identified a significant number of genes of cyanobacterial or algal origin by comparing the sequences with GenBank database using BLAST. Phylogenetic analyses indicate that some of these proteins could have been acquired from a photosynthetic endosymbiont. In addition, a putative N-terminal targeting signal has been detected in some of the proteins, suggesting that *C. cohnii* contains a plastid and that these proteins are imported into this organelle. Putative metabolic pathways that may take place in the plastid are described and compared to other plastid-containing organisms. This represents the first extensive genomic analysis of a heterotrophic dinoflagellate.

## ***Materials and Methods***

### **Strain and cultivation**

The strain of *Cryptocodinium cohnii* Seligo under study was isolated from the non-clonal ATCC #30340 culture. *C. cohnii* was grown in 50 g/L glucose, 6 g/L yeast extract, and 10% artificial seawater at 27°C and pH 6.7.

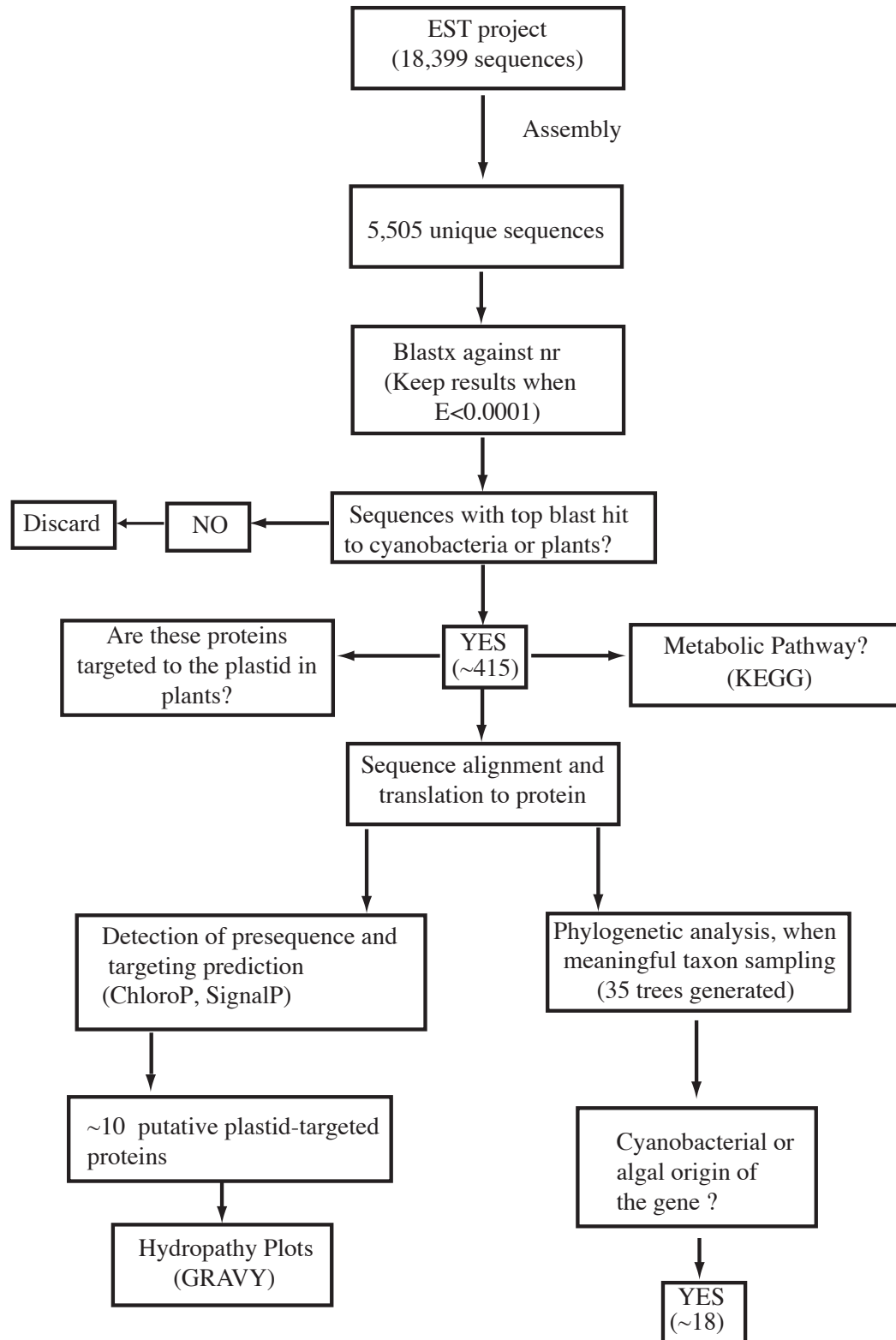
### **Library construction, sequence and analysis**

The cDNA sequences analyzed here were identified in an EST survey of *C. cohnii*, from which all genes associated with fatty acid biosynthesis have been culled. The cDNA library construction and sequencing was described elsewhere (Lippmeier et al. 2002). The sequences were assembled and the unique sequence reads or contigs were compared to the nonredundant and EST NCBI sequence databases using the BLAST tool. Only blast hits with an E-score  $< 10^{-4}$  were considered further.

### **Identification of plastid-associated genes**

A *C. cohnii* database with blast results was searched for genes whose top blast hit corresponded to plants, algae, or cyanobacteria. In addition, the function of the gene and the metabolic pathway were assessed using KEGG (<http://www.genome.jp/kegg/>). Proteins known to be involved in plastid functions were also used to directly search the *C. cohnii* database. Twenty-nine EST sequences have been deposited in GenBank under the accession numbers EB086306-EB086334.

Figure V.1. Flow diagram indicating the procedure used to identify putative plastid-derived genes and plastid-targeted proteins in the heterotrophic dinoflagellate *Cryptocodinium cohnii*.



## Phylogenetic analyses

I further analyzed the putative plastid-associated genes to infer their phylogenetic origin. Potential homologs of those genes were obtained from GenBank database and aligned manually using MacClade (Maddison and Maddison 2000). Single-gene datasets based on amino acids were analyzed with TreePuzzle 5.2 (Schmidt et al. 2002), under the JTT model of amino acid substitution (Jones et al. 1992), with eight rate categories and invariant sites estimated from the dataset. ML protein analyses were done using PhyML (Guindon and Gascuel 2003) with four gamma-distributed rate categories and invariant sites estimated from the dataset. Support for branches was obtained by bootstrapping with 100 replicates.

## Targeting prediction

The correct translation frame of each protein was inferred from blast results and alignment of the proteins was useful for detecting an N-terminal presequence in some of *C. cohnii* genes. Targeting signals were predicted using SignalP (<http://www.cbs.dtu.dk/services/SignalP/>) (Bendtsen et al. 2004), and ChloroP (<http://www.cbs.dtu.dk/services/ChloroP/>) (Emanuelsson et al. 1999). The program ChloroP predicts the presence of a chloroplast transit peptide (cTP) and peptidase cleavage site. This program has been trained with green algal sequences (Emanuelsson and Heije 2001), and thus, it may not be particularly helpful to detect transit peptides in other photosynthetic lineages, such as dinoflagellates. The program SignalP detects signal peptides, as well as signal peptidase cleavage site (Emanuelsson and Heije 2001) and it may be the most useful to detect targeting signals for secondary plastids (Obornik and Green 2005).

## ***Results***

### **Putative plastid-associated genes in *C. cohnii***

A total of 18,399 clones have been sequenced from an EST project of the dinoflagellate *C. cohnii*, representing 5,505 unique sequences. Figure V.1 indicates the steps leading to the detection of putative plastid-associated genes in *C. cohnii*. First, genes were considered likely to be plastid-associated when the top blast hit corresponded to cyanobacterial or plant gene sequences. A significant number of unique sequences (415) had their most significant similarity (based on blastx results against nr NCBI database) to *Arabidopsis*, *Oryza*, *Synechococcus*, *Nostoc*, or *Synechocystis*. From this list of genes, I identified those in which the blast results included plastid-targeted proteins in other organisms, such as plants, and I recorded the metabolic pathway to which they belong. Phylogenetic analyses of putative plastid-associated genes were performed whenever possible depending on availability of homologous sequences from other organisms. In some cases, too few homologs were available to enable meaningful phylogenetic analysis. The proteins encoded by putative plastid-associated genes in *C. cohnii* were also analyzed using two target prediction programs: ChloroP and SignalP. For some proteins, only a partial sequence was present and did not include the targeting signal; in such cases, it was not possible to predict where the sequence would be targeted. A protein was considered to be a candidate plastid-associated gene when the trees showed a close relationship with plants, algae, or cyanobacteria (see below). For those proteins with a predicted plastid-targeted signal, hydropathy plots were obtained. Overall, 18 genes are strong



Table V.1. Putative plastid-associated genes in *C. cohnii*

GenBank accession number	Putative gene product/Gene name	N-terminal extension	ChloroP p-values	SignalP p-values	EC number a	Metabolic Pathway	Alignment size
EB086308	Rubisco	incomplete	0.552	0.92	4.1.1.39	unknown	459 aa
EB086309	Delta-aminolevulinic acid dehydratase (ALADH, HemB)	complete	0.547	0.832	4.2.1.24	Heme biosynthesis	353 aa
EB086310	Protoporphyrinogen oxidase (HemG)	missing	-	-	1.3.3.4	Heme biosynthesis	246 aa
EB086311	1-deoxy-D-xylulose 5-phosphate reductoisomerase (IspC)	complete	0.53	0.59	1.1.1.267	Isoprenoid biosynthesis	246 aa
EB086312	4-Diphosphocytidyl-2C-methyl-D-erythritol synthase (IspD)	incomplete	0.559	< 0.5	2.7.760	Isoprenoid biosynthesis	246 aa
EB086306	1-hydroxy-2-methyl-2-(E)-butenyl-4-diphosphate synthase (IspG)	incomplete	0.531	0.836	1.17.4.3,	Isoprenoid biosynthesis	287 aa
EB086307	1-hydroxy-2-methyl-2-(E)-butenyl-4-diphosphate synthase (IspG)	complete	0.519	0.73	1.17.4.3,	Isoprenoid biosynthesis	287 aa
EB086315	SufB	missing	-	-		FeS cluster assembly	464 aa
EB086314	SufC	incomplete	0.578	0.836		FeS cluster assembly	231 aa
EB086316	ascorbate peroxidase	missing	-	-	1.11.1.11	Ascorbate metabolism	144 aa
EB086317	ascorbate peroxidase	missing	-	-	1.11.1.11	Ascorbate metabolism	144 aa
EB086318	Monodehydroascorbate reductase	missing	-	-	1.6.5.4	Ascorbate metabolism	321 aa
EB086321	Glutathione reductase	missing	-	-	1.8.1.7	Glutathione metabolism	506 aa
EB086322	phospholipid-hydroperoxide glutathione peroxidase	missing	-	-	1.11.1.12	Glutathione metabolism	181 aa
EB086329	Hypothetical protein 1 [At2g37660]	missing	-	-			168 aa
EB086328	Hypothetical protein 2 [At3g61320]	missing	-	-			90 aa
EB086323	Adenylate kinase	missing	-	-	2.7.4.3	Purine metabolism	179 aa
EB086324	Adenylate kinase	missing	-	-	2.7.4.3	Purine metabolism	179 aa
EB086320	Nitrate transporter	complete	< 0.5	0.713			345 aa
EB086319	Nitrate transporter	missing	-	-			345 aa
EB086326	Branched-chain amino acid aminotransferase 5	missing	-	-	2.6.1.42	Valine, leucine, isoleucine metabolism	239aa
EB086327	Branched-chain amino acid aminotransferase 5	incomplete	0.553	<0.5	2.6.1.42	Valine, leucine, isoleucine metabolism	239aa
EB086325	Branched-chain amino acid aminotransferase-like protein 3	missing	-	-	2.6.1.42	Valine, leucine, isoleucine metabolism	278 aa
EB086313	Farnesyl pyrophosphate synthetase (FPP synthetase) (FPPS)	incomplete	0.546	< 0.5	2.5.1.10	Carotenoid biosynthesis	351 aa

a- Enzyme Commission (EC) number

candidates for plastid-associated genes, and ten are predicted to be localized to the plastid (Table V.1).

### **Phylogenetic analyses**

A gene encoding rubisco (ribulose-1,5-bisphosphate carboxylase/oxygenase) was identified in *C. cohnii*. Phylogenetic analyses show a close relationship to peridinin-containing dinoflagellate form II rubisco, which shares a common ancestor with alpha-proteobacteria (Figure V.2). Two other well-supported clades correspond to form I rubisco, including green algal and cyanobacterial form Ia and red algal (plus organisms with red algal-derived plastids) form Ib rubisco. A plastid-targeting signal consisting of a signal peptide and a transit peptide was predicted for *C. cohnii* rubisco (Table V.1).

Individual phylogenetic analyses are, in general, not fully resolved due to the small alignment size (Table V.1). Eight genes from *C. cohnii* show moderate to strong support for a relationship with other algae (red or green algae, including plants) and/or cyanobacteria (Figures V.2-V.4). Phylogenetic analyses of the genes *hemB*, *hemG*, glutathione reductase, and monodehydroascorbate reductase from *C. cohnii* show a relationship to green algal (or plant) genes (Figures V.2, V.3). Other genes (FPPS, ascorbate peroxidase, branched-chain aa aminotransferase 5, glutathione peroxidase) are related to red or chl c containing algae with variable support (Figures V.4, V.5). Genes encoding nitrate transporter and two hypothetical proteins (EB086328-9), as well as the genes *IspC*, *IspD*, *IspG*, *sufB*, and *sufC*, cluster with algal and cyanobacterial sequences but do not show a strong affinity for either green or red algae (Figures V.3-V.5). In the case of branched-chain aa

Figure V.2. Phylogenetic analyses of putative plastid-targeted genes in *C. cohnii*, plotted on a common scale. Best trees based on proteins under maximum likelihood (ML) using the program PhyML. Numbers on branches correspond to bootstrap values (above) from the PhyML analyses, and quartet puzzling support values (below) from TreePuzzle analyses, and are shown when > 50. Predicted localization of proteins is indicated when known. P= plastid-targeted nuclear-encoded, Pe= plastid encoded.

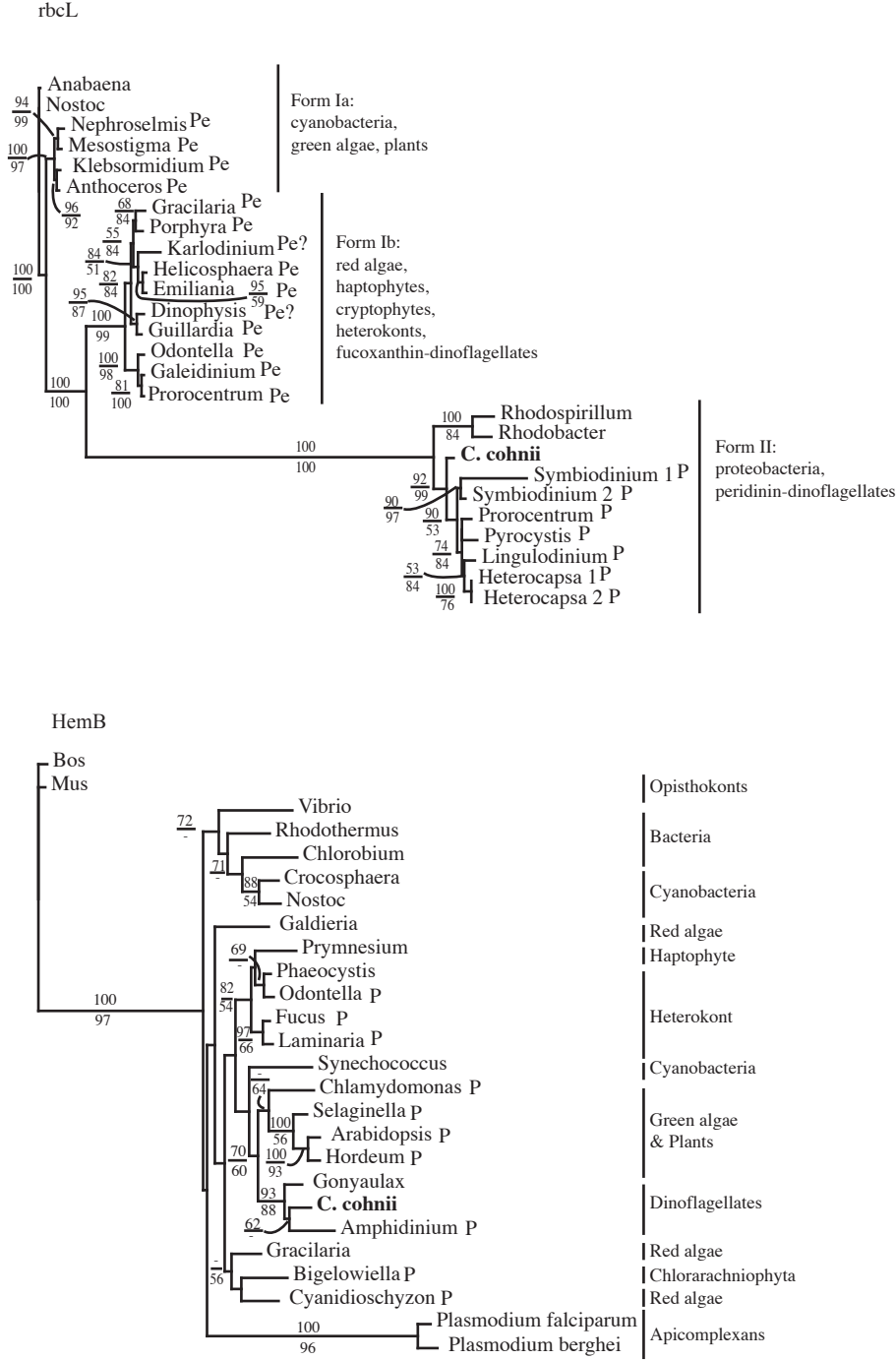


Figure V.3. Phylogenetic analyses of putative plastid-associated genes in *C. cohnii*, plotted on a common scale. Best trees based on proteins under maximum likelihood (ML) using the program PhyML. Numbers on branches correspond to bootstrap values (above) from the PhyML analyses, and quartet puzzling support values (below) from TreePuzzle analyses, and are shown when > 50. Predicted localization of proteins is indicated when known. P= plastid-targeted nuclear-encoded, M= mitochondria-targeted, C= cytosolic protein, Pe= plastid encoded.

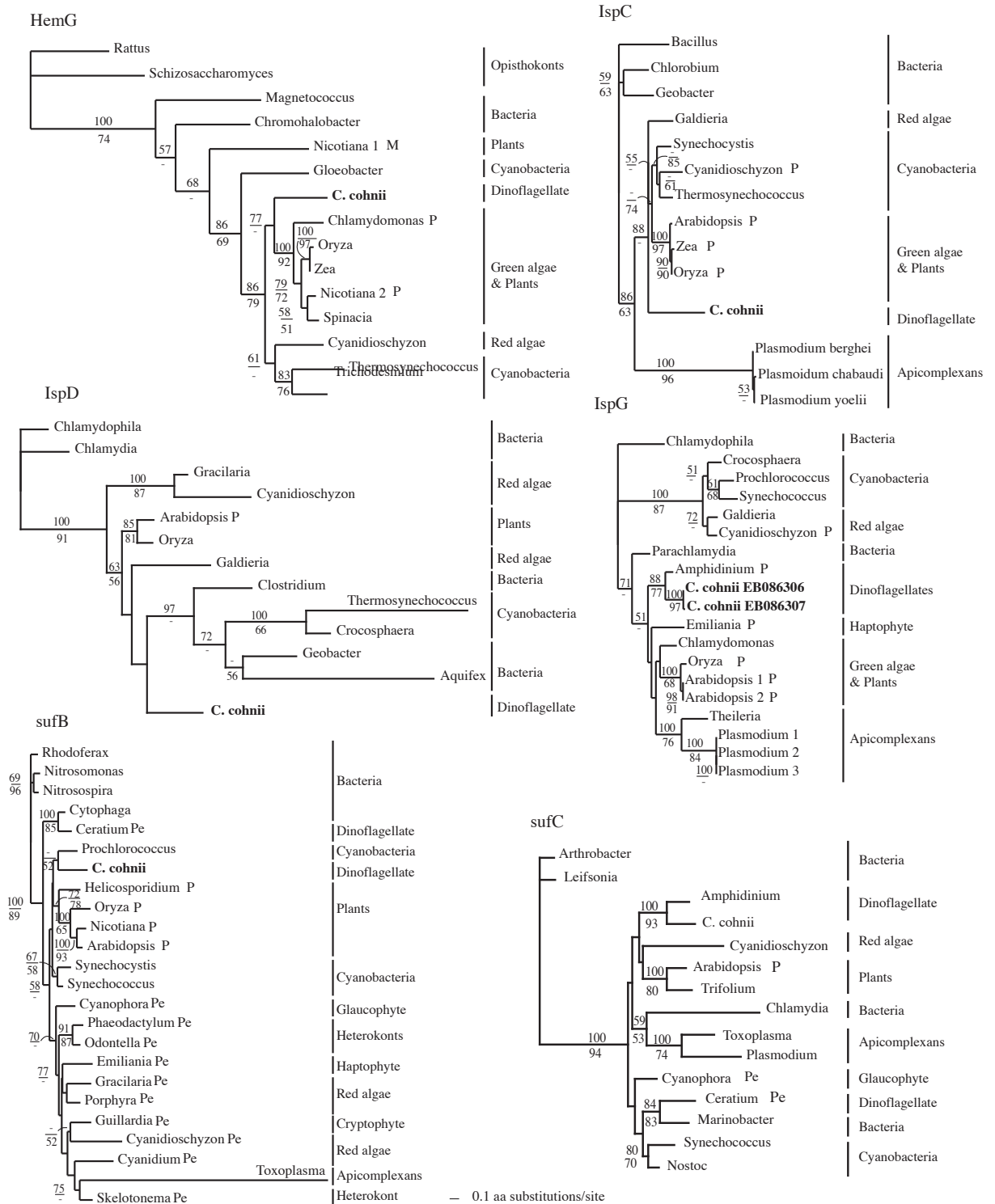


Figure V.4. Phylogenetic analyses of putative plastid-associated genes in *C. cohnii*, plotted on a common scale. Best trees based on proteins under maximum likelihood (ML) using the program PhyML. Numbers on branches correspond to bootstrap values (above) from the PhyML analyses, and quartet puzzling support values (below) from TreePuzzle analyses, and are shown when > 50. Predicted localization of proteins is indicated when known. P= plastid-targeted nuclear-encoded, M= mitochondria-targeted, C= cytosolic protein.

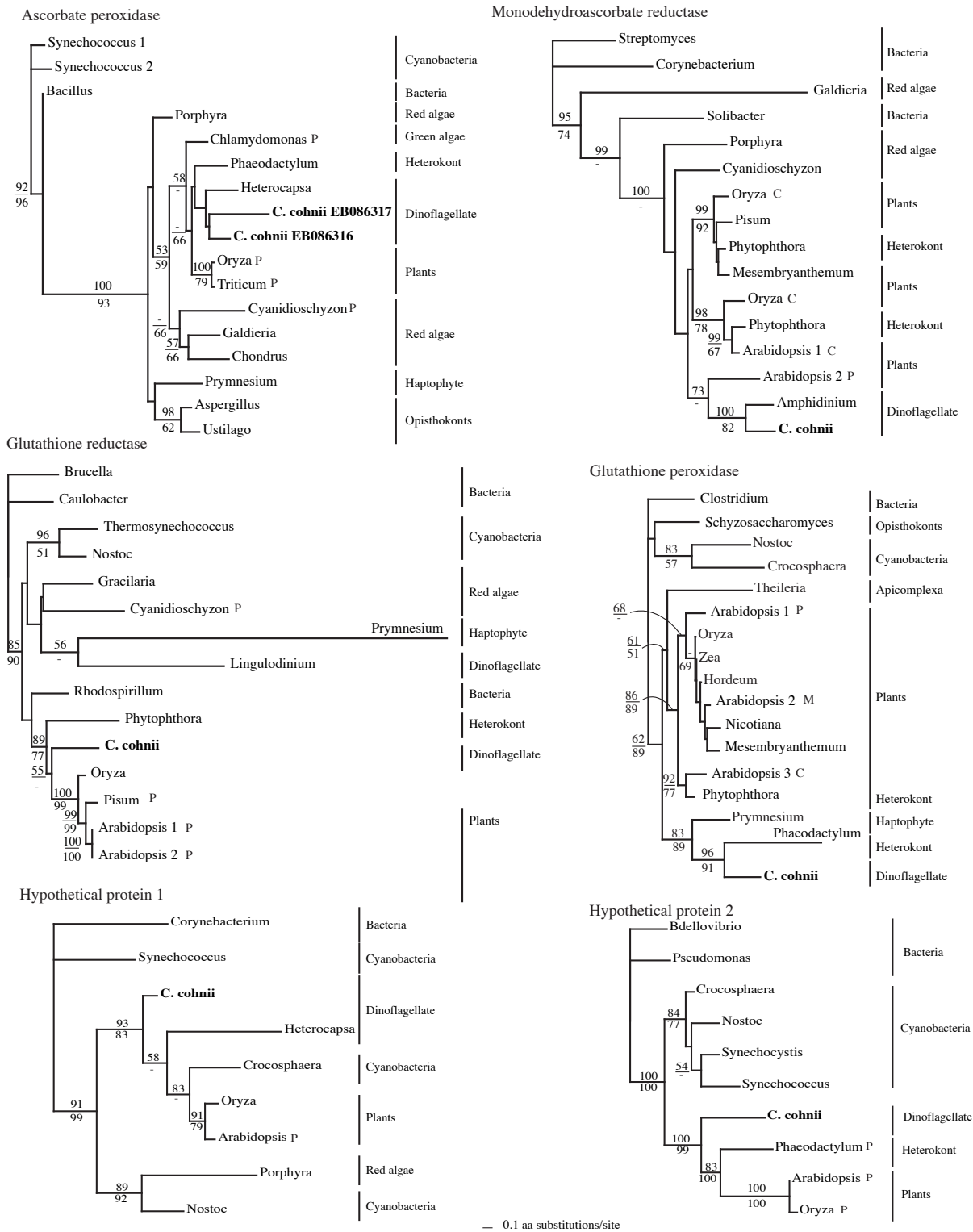
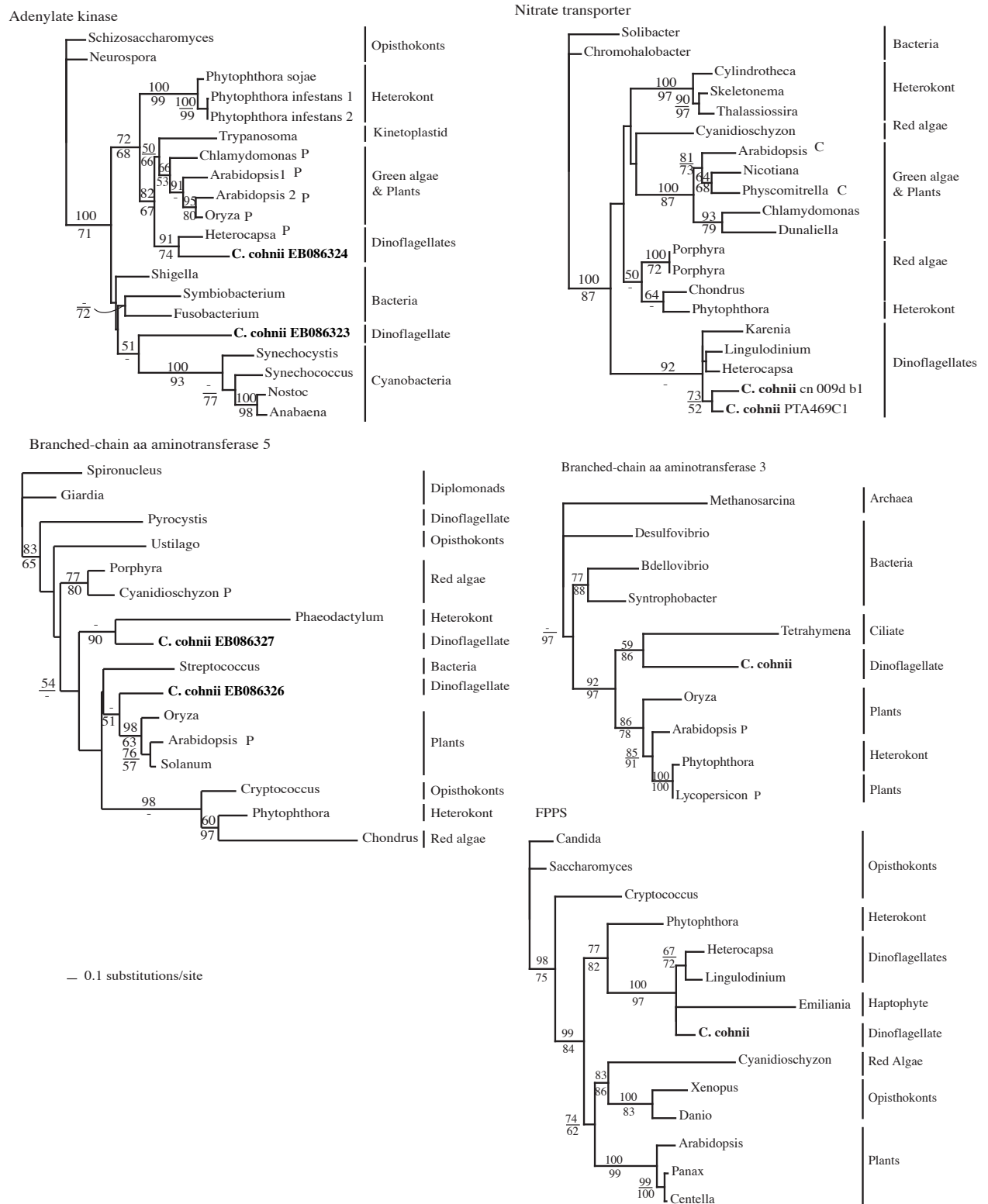


Figure V.5. Phylogenetic analyses of putative plastid-associated genes in *C. cohnii*, plotted on a common scale. Best trees based on proteins under maximum likelihood (ML) using the program PhyML. Numbers on branches correspond to bootstrap values (above) from the PhyML analyses, and quartet puzzling support values (below) from TreePuzzle analyses, and are shown when > 50. Predicted localization of proteins is indicated when known. P= plastid-targeted, C= cytosolic protein.

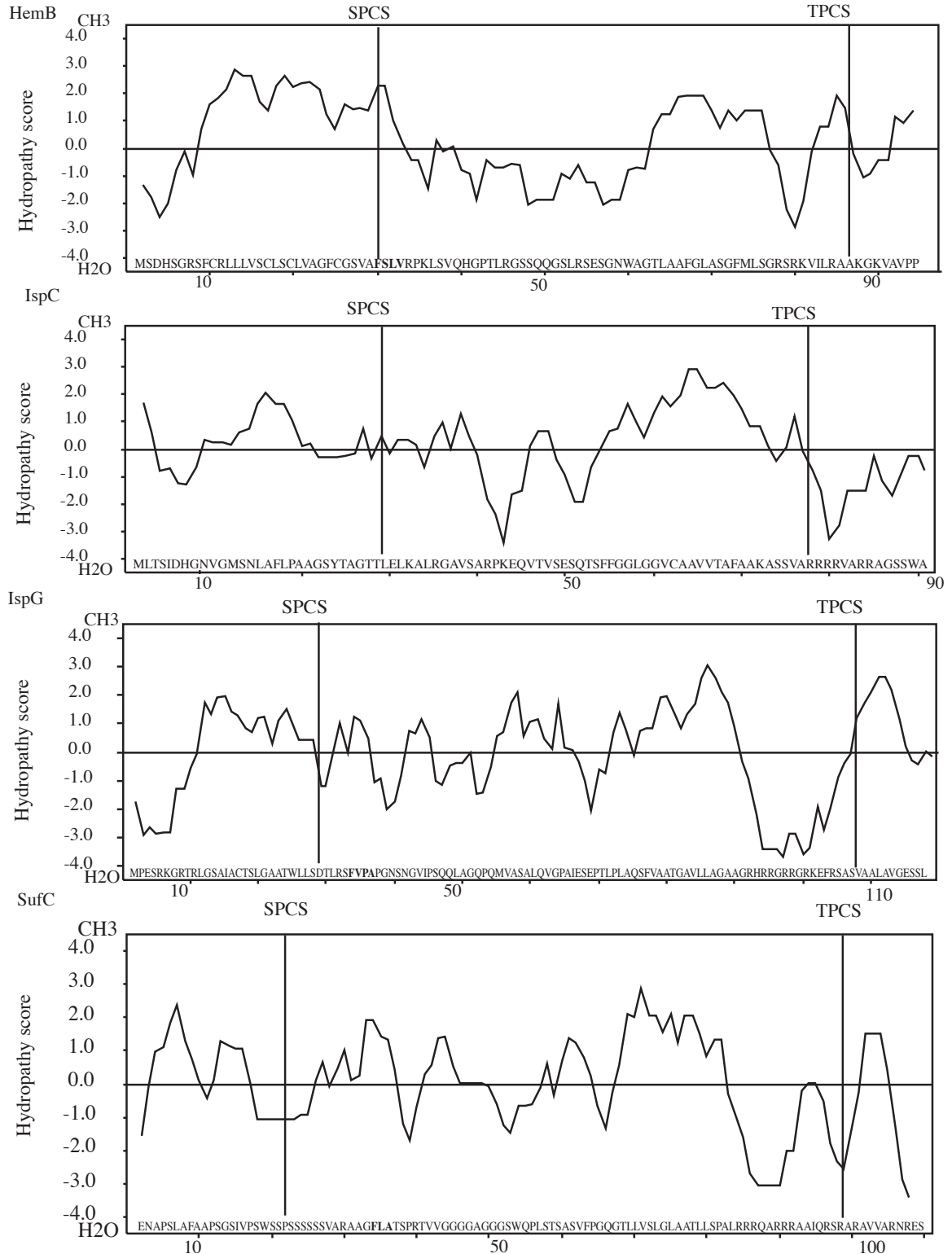


aminotransferase-like protein 3 and adenylate kinase, no red or chl c containing algae other than dinoflagellates were present in the analysis; *C. cohnii* clusters with green algae and land plants (Figure V.5).

### Targeting signal prediction

Putative plastid-targeting signals are detected in several genes: *rbcL*, *hemB*, *IspC*, *IspD*, *IspG*, *sufC*, FPPS, nitrate transporter, and branched-chain aa aminotransferase 5 (Table V.1). For some genes (*IspD*, FPPS, and branched-chain aa aminotransferase 5) a chloroplast target peptide is predicted by ChloroP and SignalP does not detect a signal peptide. Only a partial N-terminal sequence is available for those genes, and thus, the signal peptide could be missing from the library but present in the full-length protein. In contrast, the gene encoding for a nitrate transporter presents a signal peptide, but no transit peptide is detected. It is possible that this protein is localized to the plastid outer membrane and in this case a transit peptide would not be required (Li and Chen 1996), although a particular domain may be present for integration in the membrane. A number of genes from *C. cohnii* show a close relationship to genes in photosynthetic eukaryotes, including genes encoding plastid-targeted proteins in plants, but the EST sequence is not complete and thus, the presence of a targeting signal could not be tested (Table V.1). Both a signal and a transit peptide are predicted for six genes (Table V.1). Hydropathy plots from genes with complete N-terminal sequences (*HemB*, *IspC*, *IspG*) and from the partial presequence from *sufC* are shown in Figure V.6. The signal peptide is characterized by a high content of hydrophobic amino acids. Signal peptidase cleavage site is in some cases followed by a phenylalanine-containing motif, including three other

Figure V.6. Kyte-Doolittle hydropathy plots, with window size of 5 amino acids. The putative signal peptidase (SPCS) and transit peptidase (TPCS) cleavage sites are indicated with a vertical line. Each graph corresponds to the gene sequence shown below it.





hydrophobic residues. The remainder of the transit peptide is often enriched in hydroxylated amino acids, such as serine and threonine. These four proteins (*HemB*, *IspC*, *IspG*, *sufC*) contain a second hydrophobic region towards the C-terminal end of the transit peptide. This region is sometimes followed by an arginine-rich region (Figure V.6).

## ***Discussion***

The generation of a cDNA library from the heterotrophic dinoflagellate *Cryptothecodinium cohnii* is a powerful source for functional and phylogenetic comparisons of expressed genes with other organisms, as well as for increasing our understanding of dinoflagellate evolution. This is the first extensive genomic study done on a heterotrophic dinoflagellate and the first detailed analysis of the plastid-related metabolic pathways investigating this lineage. I identified a significant number of plastid-associated genes expressed in *C. cohnii* by blasting the sequences to GenBank nonredundant database, performing phylogenetic analyses to infer the evolutionary relationships, and finally detecting plastid-targeting signals. The data presented here suggest that *C. cohnii* evolved from a plastid-bearing ancestor, and may possess a reduced plastid where a number of metabolic pathways still take place.

## **Evidence for past endosymbiosis**

The presence of eighteen genes with a cyanobacterial origin in the nuclear genome of *C. cohnii* argues for an earlier endosymbiotic event, in which those genes were transferred from the endosymbiont to the host nucleus. An alternative scenario would be a lateral gene transfer of these plastid-associated genes from a

cyanobacterium or a photosynthetic eukaryote to *C. cohnii*. However, this hypothesis seems unlikely given the number of horizontal gene transfers that would be required.

The mevalonate-independent pathway, DXP or MEP pathway, for terpenoid biosynthesis is only known for bacteria and plastid-containing eukaryotes. The identification in *C. cohnii* of three enzymes involved exclusively in this pathway represents strong evidence for a cyanobacterial origin of those genes and early acquisition of plastids in dinoflagellates. Horizontal gene transfer is not particularly well demonstrated in eukaryotes; thus, the most likely origin for genes involved in the DXP pathway is through endosymbiosis. Furthermore, the gene *sufB*, which is plastid-associated in all known eukaryotes, is present in *C. cohnii*. In addition, genes encoding enzymes involved in plastid-related metabolism in plants and algae, such as heme biosynthesis and iron-sulfur cluster assembly, are present in *C. cohnii* and probably acquired from a photosynthetic endosymbiont.

## **Rubisco**

Genes related to photosynthesis are usually lost or become nonfunctional pseudogenes in heterotrophic or parasitic organisms, even if the organelle is maintained (Gockel and Hachtel 2000; Cai et al. 2003; Foth and McFadden 2003). The gene *rbcL* is an exception because it is retained intact in the plastid genome of various organisms that lost their photosynthetic ability (Wolfe and dePamphilis 1998; Sekiguchi et al. 2002). *C. cohnii* has maintained the gene *rbcL* in its nuclear genome and it is transcribed. The enzyme Rubisco exists in two different forms: form I consists of 8 large and 8 small subunits; form II consists of only two large subunits (Watson and Tabita 1997). Different types of form I Rubisco are present in green

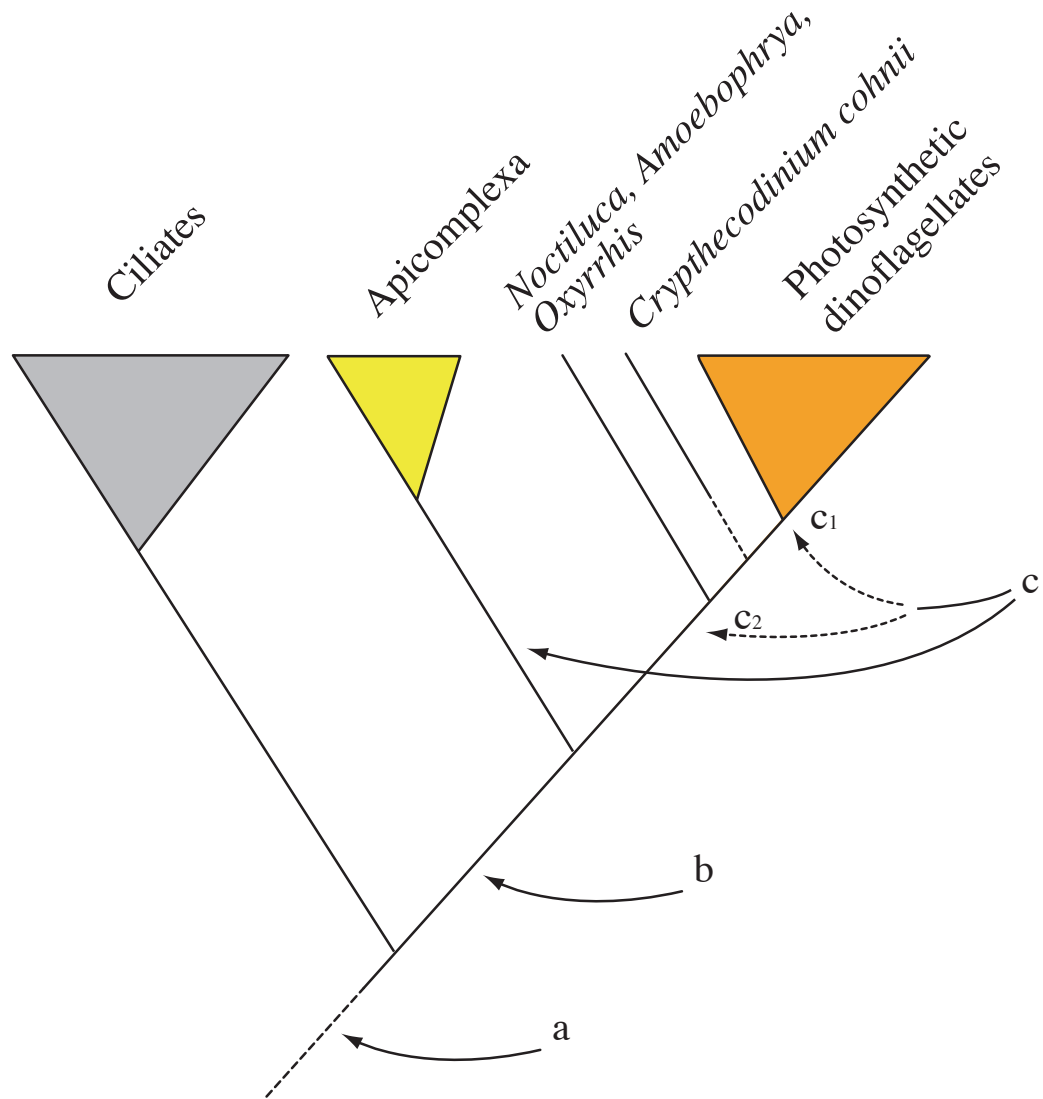
algae, red algae, chl c algae, cyanobacteria, and some alpha, beta or gamma proteobacteria, while form II Rubisco has been found in alpha, beta and gamma purple bacteria, and peridinin-containing dinoflagellates (Morse et al. 1995). The gene *rbcL* from *C. cohnii* shares an evolutionary history with form II Rubisco genes from peridinin-containing dinoflagellates. The evolutionary origin of this protein in dinoflagellates is not related to the plastid endosymbiont, but laterally transferred from an alpha-proteobacterium (Morse et al. 1995; Delwiche and Palmer 1996). The presence of this gene in *C. cohnii* suggests that this horizontal gene transfer occurred early in the evolution of dinoflagellates. Non-photosynthetic organisms (e.g. euglenoids, hemi- and holoparasitic plants) often contain the gene *rbcL* as a pseudogene or intact open reading frame, but whether the protein is functional or not remains unclear. An almost complete sequence from this gene in *C. cohnii* suggests that the gene is conserved but it is unknown whether the protein is translated and functional or not. In photosynthetic eukaryotes, the gene *rbcL* is involved mainly in carbon fixation; but it is not clear why this gene is conserved in non-photosynthetic lineages, such as *C. cohnii*. Possible explanations include that it may act as oxygenase, be involved in glycine and serine biosynthesis, perform limited carbon fixation, or have a yet unidentified function (Wolfe and dePamphilis 1998; Sekiguchi et al. 2002).

### **Models of evolution**

Dinoflagellates, together with cryptophytes, haptophytes, and heterokonts, contain secondary plastids with chlorophyll c as a main photosynthetic pigment. Chl c containing plastids descend from a common red algal ancestor (Yoon et al. 2002b;

Bachvaroff et al. 2005), but the relationships among the host cells are not well understood. At least two models of host cell evolution have been postulated that are consistent with the monophyly of the chl c containing plastids. One model states that the chl c containing host cells (collectively called chromalveolates) descend from a common ancestor and that they acquired the plastid from a red alga before the divergence of the four lineages (Cavalier-Smith 1981; Cavalier-Smith 2002). Under this model, the non-photosynthetic organisms that belong to these lineages are secondarily heterotrophic and plastid-associated genes are expected to exist in their nuclear genome. An alternative model suggests that chl c containing host cells are not closely related and that these lineages acquired their plastids in independent endosymbiotic events (from each other or the red algae) (Cavalier-Smith et al. 1994; Bachvaroff et al. 2005). In this case, basal heterotrophic members of these groups may have never contained a plastid and no plastid-associated genes are expected in the nuclear genome. If we focus on the alveolates, which include dinoflagellates, apicomplexans, and ciliates, we could list three different patterns of plastid acquisition (Figure V.7): a- plastids were acquired before the divergence of ciliates, apicomplexans, and dinoflagellates; b- plastids were acquired after the divergence of ciliates and before the split of apicomplexans and dinoflagellates; c- plastids were acquired independently in apicomplexans and dinoflagellates. Under hypotheses a and b, all dinoflagellates (including heterotrophic and aplastidic ones) may have plastid-associated genes in their nuclear genome remnant of an earlier endosymbiotic event, and apicoplasts and peridinin-containing dinoflagellate plastids could be closely related. Under hypothesis c, basal heterotrophic dinoflagellates (such as *C.*

Fig V.7. Models of plastid acquisition (a-c) in Alveolates. Ciliates and the dinoflagellates *Noctiluca*, *Amoebophrya*, *Oxyrrhis*, and *Crypthecodinium* are heterotrophic and no plastid is known. Apicomplexans contain a nonphotosynthetic apicoplast. Clade with photosynthetic dinoflagellates includes dinoflagellates hosting different type of plastids and secondarily heterotrophic species.



*cohnii*) may have never contained a plastid (Figure V.7, c1), and apicoplasts and peridinin-plastids may not be related. Analysis of nuclear genomes of basal heterotrophic chromalveolates would be useful to distinguish among these hypotheses and elucidate the evolutionary history of these important groups of algae. In the present study, I show that the dinoflagellate *C. cohnii* contains genes likely derived from a photosynthetic endosymbiont, suggesting an earlier plastid acquisition in dinoflagellates. The endosymbiotic event could have occurred as early as the chromalveolate hypothesis postulates, before the divergence of alveolates, heterokonts, haptophytes, and cryptophytes.

Surprisingly, I also identified some genes (monodehydroascorbate reductase, glutathione reductase, glutathione peroxidase, adenylate kinase, branched-chain aa aminotransferase 3) from a heterotrophic heterokont (the oomycete *Phytophthora*) that show a close relationship with genes from photosynthetic eukaryotes (Figures V.4, V.5). These genes may have been acquired from a photosynthetic endosymbiont, pushing back the estimated event of plastid acquisition in chromalveolates. However, more detailed studies are necessary to eliminate other possible scenarios.

### **Origin of plastid-derived genes in *C. cohnii***

The plastids from peridinin-containing dinoflagellates are derived from red algae and thus, phylogenetic analyses of plastid genes show a close relationship to red algae or lineages with red algal-derived plastids (Yoon et al. 2002a; Bachvaroff et al. 2005). A red algal ancestry is expected for plastid-associated genes in *C. cohnii*; however not all plastid gene trees show this. In most cases, the relationships of *C. cohnii* genes are not resolved and the best tree shows a close relationship with green

algae (including plants), red algae, or chl c containing algae. One reason for the lack of support for a red algal origin in some gene trees could be the reduced taxon sampling; in most cases, analyses include one red alga (*Cyanidioschyzon merolae*), or only a few chl c containing algae. Sequences from the red alga *C. merolae* are very divergent compared to other organisms, probably due to the extreme conditions in which it lives (45°C and pH 1.5) (Matsuzaki et al. 2004). A special case is the gene *hemB*, which may be the result of a lateral gene transfer from a green alga to dinoflagellates (Hackett et al. 2004). Because of its independent origin, *hemB*, like *rbcL*, does not represent evidence of past endosymbiosis.

### **Predicting protein targeting**

With strong evidence for genes derived from a photosynthetic endosymbiont in *C. cohnii*, I was intrigued by the localization of the encoded proteins in the cell. Proteins formed in the cytosol that function in other cellular compartments need to be recognized and transported across membranes by specific translocators. In the case of plastid-targeted proteins, the N-terminal targeting signals are recognized based on characteristics of the amino acids: hydrophobicity and secondary structure (Kroth 2002). Plastid-targeted proteins in organisms with primary plastids (surrounded by only 2 membranes), such as green algae, land plants and red algae, contain a chloroplast transit peptide, which is low in acidic residues but with high content of hydroxylated residues (Emanuelsson et al. 2000; Kroth 2002). In organisms with secondary plastids, which are surrounded by 3 or 4 membranes and located in the endomembrane system of the host, the targeting presequence consists of two parts: a signal and a transit peptide, indicating that proteins are first transported into the

secretory pathway, and then directed to the plastid (van Dooren et al. 2001; Kroth 2002). In peridinin-containing dinoflagellates, the N-terminus of the transit peptide contains a short motif that consists of a phenylalanine residue followed by three hydrophobic ones (Patron et al. 2005). A similar conserved motif has been described for heterokonts (Kilian and Kroth 2005), but has only been recognized in a small number of plastid-targeted proteins in haptophytes and fucoxanthin-containing dinoflagellates (Patron et al. 2006).

Alignments of some proteins from *C. cohnii* revealed the presence of an N-terminal presequence (in comparison to bacterial or cytosolic homologs) predicted to contain a signal peptide and a transit peptide adjacent to each other, suggesting that these proteins were targeted to a secondary plastid rather than staying in the cytosol. A number of plastid-associated proteins from *C. cohnii* share characteristics described for peridinin-containing dinoflagellates, including a phenylalanine-containing motif following the signal peptidase cleavage site. It has been noted that some proteins from peridinin-containing dinoflagellates contain a second hydrophobic region, thought to be important in the processing of the protein, towards the C-terminal end of the transit peptide followed by an arginine-rich region (Patron et al. 2005). This pattern was observed in targeting signals from *hemB*, *IspC*, *IspG*, and *sufC* in *C. cohnii*.

Prediction of plastid-targeting signals in *C. cohnii* proteins, as well as the presence of proteins that are either plastid-encoded or plastid-targeted in all plastid-containing eukaryotes (*IspC*, *IspD*, *IspG*, *sufB*), lead me to think that a reduced plastid could be physically present in *C. cohnii*. An early ultrastructural study done on this species described the presence of atypical plastids (Kubai and Ris 1969). Using



transmission electron microscopy from another isolate of *C. cohnii* (ATCC #30340), I could not identify a structure that clearly resembles a plastid, although unknown membrane structures were observed (data not shown). Non-photosynthetic plastids are not easy to recognize and they have at times first been identified by the presence of plastid-associated genes (Gardner et al. 1991; de Koning and Keeling 2004).

Conservation of plastid targeting signals is also good evidence for the presence of a plastid and proteins transport into the organelle. If plastids were lost, maintenance of plastid targeting signals would affect the localization and function of these proteins, and therefore either the targeting signals (or the protein altogether) would be expected to be lost or degenerated. Assuming that *C. cohnii* contains a reduced plastid, it remains unknown whether it maintains a plastid genome. The gene *sufB* is plastid-encoded in all eukaryotes containing red algal-derived plastids, including red algae. The *C. cohnii* *sufB* sequence is missing the N-terminal portion, but evidence from the higher GC% content (51%) compared to minicircle genes in other dinoflagellates, and the presence of a poly-A tail suggests that *sufB* is nuclear-encoded in *C. cohnii*.

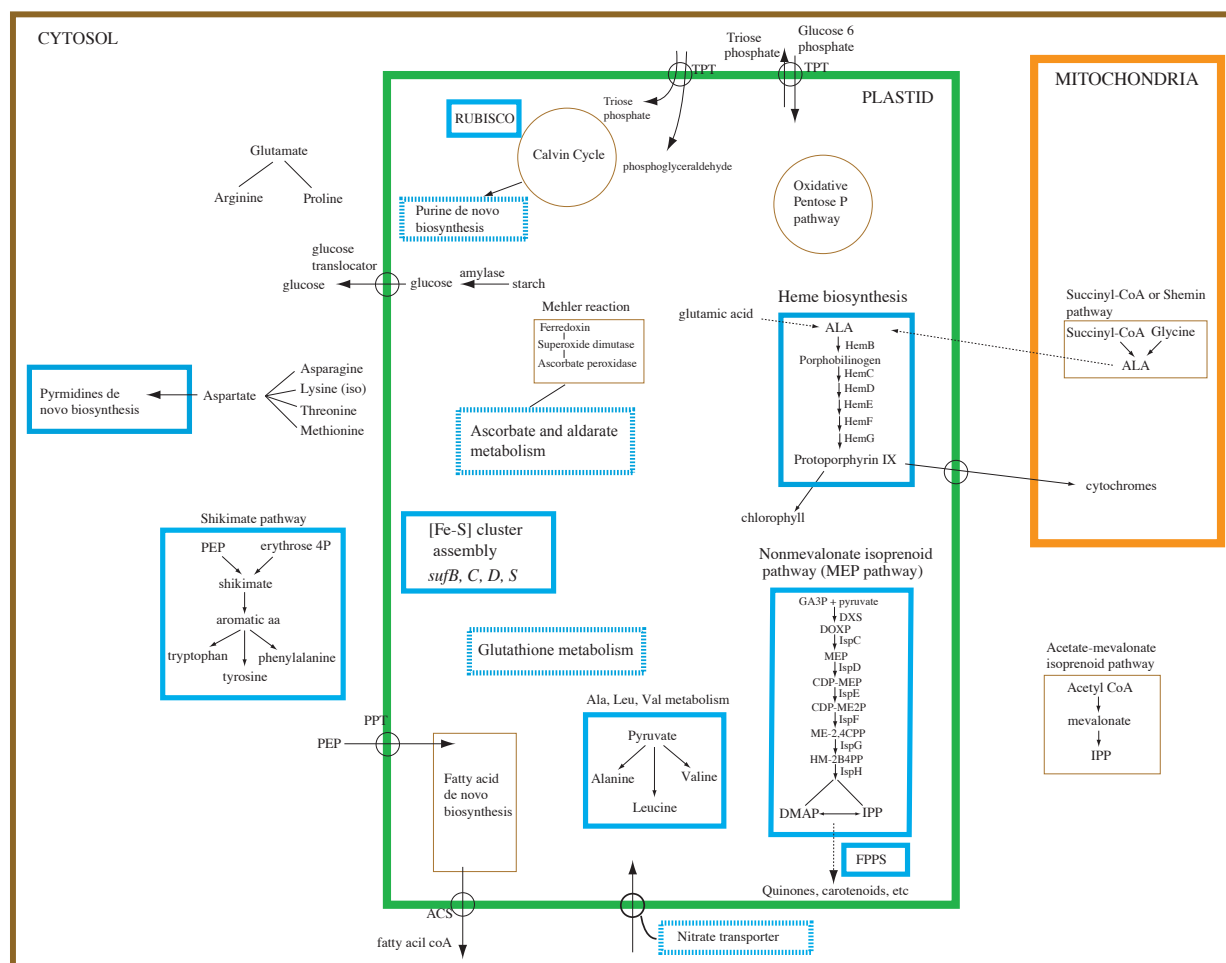
### **Plastid-related metabolism in *C. cohnii* and comparison to other algae**

The metabolism of nonphotosynthetic plastids has been only recently studied in two achlorophyllic green algae (de Koning and Keeling 2004; Borza et al. 2005) and the apicomplexan parasites (Ralph et al. 2004; Wilson 2005). Much is known about plastid-related pathways of the sister taxa to dinoflagellates; however, we are just starting to learn about the genetics and biochemistry of these fascinating organisms. No genomic studies have been performed on the metabolism of

nonphotosynthetic dinoflagellates, and very little has been done in photosynthetic ones (Bachvaroff et al. 2004; Tanikawa et al. 2004; Hackett et al. 2005; Lidie et al. 2005). Here, I present a description of the main metabolic pathways related to plastid functions and localization in the heterotrophic dinoflagellate *C. cohnii* (Figure V.8) as inferred from genes identified in this EST project.

Heme biosynthesis is performed by prokaryotes and eukaryotes (photosynthetic or not) and is required for both cytochrome and chlorophyll biosynthesis (Figure V.8). This pathway takes place in the plastid of photosynthetic eukaryotes (Cornah et al. 2003) and is catalyzed by enzymes of mosaic origin: proteins derived from the plastid and host cell (Obornik and Green 2005). Two genes have been identified in the EST survey in *C. cohnii* involved in this metabolic pathway:  $\delta$ -aminolevulinic acid dehydratase (ALADH, or *hemB*), and protoporphyrinogen oxidase (*hemG*). The enzyme *hemB* converts ALA to porphobilinogen. Reactions starting with ALA through the formation of coproporphyrinogen III occur in the plastid of photosynthetic eukaryotes (Cornah et al. 2003), and in the apicoplast of the apicomplexan *P. falciparum* (Ralph et al. 2004). Later, coproporphyrinogen III is converted to protoporphyrinogen IX and then, into protoporphyrin IX by *hemG*. The last two reactions take place in the plastids of photosynthetic eukaryotes, and in the cytosol or mitochondria of *P. falciparum* (Ralph et al. 2004). Both genes identified in *C. cohnii* (*hemB*, *hemG*) are derived from a photosynthetic endosymbiont and putatively plastid-targeted, suggesting that heme biosynthesis occurs in the organelle, probably followed by the transport of protoporphyrin IX to the mitochondria for cytochrome synthesis.

Figure V.8. Plastid-related metabolism. Pathways identified in *C. cohnii* are framed by thicker lines. Localization of pathways in different compartments is based on targeting signal of proteins in *C. cohnii*. Dashed lines indicate the presence of enzymes from the corresponding pathway in *C. cohnii*, which are derived from a photosynthetic ancestor and probably localized to the plastid, but the presence is missing. Fine lines surround other known pathways not identified in *C. cohnii*.



Terpenoids are naturally occurring organic products, composed of a number of isopentenyl diphosphate (IPP) units. The precursor IPP and its isomer DMAP (dimethylallyl diphosphate) can be produced by two different pathways: the classical acetate/mevalonate (MVA) pathway and the alternative non-mevalonate (DXP/MEP) pathway (Buchanan et al. 2000; Lange et al. 2000). The latter pathway has only been described in bacteria and plastids of eukaryotes (Rohmer et al. 1993; Lichtenthaler et al. 1997), while the MVA pathway takes place in the cytosol of animals and fungi. Some plants, and algae contain genes for both pathways; MVA pathway takes place in cytosol and the alternative one occurs in the plastid (Disch et al. 1998; Cunningham et al. 2000). Other algae synthesize all isoprenoids using only one pathway (Disch et al. 1998). Three enzymes involved in terpenoid biosynthesis have been identified from the EST library of *C. cohnii*: DXP reductoisomerase (*IspC*), 4-diphosphocytidyl-2C-methyl-D-erythritol synthase (*IspD*), and 2-hydroxy-2-methyl-2-(E)-butenyl-4-diphosphate synthase (*IspG*), indicating the presence of the non-mevalonate pathway (Figure V.8). All of these enzymes seem to be targeted to the apicoplast in apicomplexans (Ralph et al. 2004). In the dinoflagellate *C. cohnii*, the three enzymes are probably plastid-localized, as indicated by the presence of an N-terminal targeting signal. Enzymes from the cytosolic mevalonate pathway were not found in the *C. cohnii* library.

One enzyme, farnesyl diphosphate synthase (FPPS), involved in the synthesis of isoprenoid-end products, has been identified in *C. cohnii*. This enzyme is the starting point of a number of different pathways leading to a variety of products, such as carotenoids, quinines, prenylated proteins, etc. The *C. cohnii* gene that encodes a

putative FPPS has a close relationship to plants and a transit peptide is predicted. Homologs from plants (FPPS2) and the red alga *C. merolae* do not have an N-terminal extension, although another gene encoding FPPS (FPPS1) is plastid-targeted in plants (Cunillera et al. 2000). Carotenoid biosynthesis begins with the formation of geranyl diphosphate (GPP) from isoprenoid units and synthesis of farnesyl diphosphate (FPP) from GPP and IPP catalyzed by FPPS (Cunningham and Gantt 1998). Enzymes involved in the synthesis of carotenoids are localized to the plastid in plants and the heterokont *Thalassiosira pseudonana* (Armbrust et al. 2004) and have not been identified in the EST project of *C. cohnii*.

Another metabolic pathway that takes place in plastids is iron-sulphur cluster assembly. Iron-sulfur (Fe-S) clusters are important cofactors of Fe-S proteins (e.g. *IspG*), which are involved in numerous vital biological processes in all organisms studied. Biogenesis of Fe-S clusters requires the mobilization of sulphur and iron and it takes place in both plastids and mitochondria (Merchant 2006). Sulphur is probably derived from cysteine in the plastids through the action of cysteine desulfurase. Genes homologous to SUF components in bacteria are also found in plastid-containing eukaryotes (Wilson 2005; Xu et al. 2005; Merchant 2006). In plastids of plants and bacteria, *sufC* associates with *sufB* and *sufD* to form a complex that is required to repair labile Fe-S clusters damaged during oxidative stress and *sufS* constitutes one subunit of cysteine desulfurase (Xu et al. 2005; Merchant 2006). Two genes involved in Fe-S cluster assembly were identified in *C. cohnii*, namely *sufB* and *sufC*. Only the N-terminal sequence of the gene *sufC* was available. A putative targeting signal was

detected that directs the protein to the plastid, suggesting that this pathway takes place in the organelle.

Nucleotide biosynthesis takes place in the plastid or cytosol of photosynthetic eukaryotes. *C. cohnii* is capable of de novo purine and pyrimidine biosynthesis, as suggested by the presence of enzymes involved in these pathways. N-terminal targeting signals were missing from some sequences, including all purine synthesis-related proteins. The absence of an N-terminal extension in one gene encoding orotate phosphoribosyltransferase involved in pyrimidine biosynthesis indicates that this pathway occurs in the cytosol of *C. cohnii*, as it does in the heterokont *Thalassiosira pseudonana* (Armbrust et al. 2004).

Synthesis of amino acids takes place in the plastid of two parasitic green algae (de Koning and Keeling 2004; Borza et al. 2005), but not in the apicoplast of *Plasmodium* (Ralph et al. 2004). A few transcripts encoding enzymes involved in biosynthesis of several amino acids were identified in *C. cohnii*, but the location of these pathways could not be assessed due to the limited amount of data. An enzyme (chorismate synthase) from the shikimate pathway leading to the synthesis of aromatic amino acids was identified in *C. cohnii* and it is probably localized to the cytosol, as indicated by the absence of an N-terminal extension (data not shown). This pathway occurs in the plastid of plants and in the cytosol of fungi and apicomplexan parasites. Cytosolic localization of chorismate synthase in *C. cohnii* is congruent with a close relationship of dinoflagellate (including *C. cohnii*) and apicomplexan sequences with fungal genes (Keeling et al. 1999).

## Histones

The nucleus of dinoflagellates has many peculiar features, including permanently condensed chromosomes, a large quantity of DNA, and the absence of nucleosome structures and histones (Rizzo 1987). Dinoflagellate nuclear DNA is associated with bacterial histone-like proteins (Wong et al. 2003). Histones are the principal structural protein of eukaryotic chromosomes and are known to occur in the sister groups to dinoflagellates: ciliates and apicomplexans. There are four main groups of eukaryotic histones: H1, H2, H3, and H4. The H1 histones are larger (ca 220aac) and less conserved (only conserved in the globular central region) than the other histones. Histone H4 is one of the most conserved proteins across eukaryotes because this proteins is in contact with the other histone proteins and almost any amino acid substitution would affect its function (Thatcher and Gorovsky 1994). Two subtypes of histone H2 are known: H2A and H2B. In addition, variants of histone H2A include H2A.X and H2A.Z; they are present in nearly all eukaryotes. Histone H2A.X contains an extended N-terminal and differs from the canonical H2A in an additional conserved motif SQ(E/D) $\Phi$  ( $\Phi$  refers to a hydrophobic residue) in the C-terminal extension. The conserved motif in histone H2A.X is involved in chromatin compaction and DNA repair (Malik and Henikoff 2003). Recent genomic studies identified a putative histone H2A.X in *Alexandrium tamarense* and histone H3 in *Pyrocystis lunula* (Okamoto and Hastings 2003; Hackett et al. 2005). In this study, I identified two histone sequences in the *C. cohnii* database, increasing the number and types of histones found in dinoflagellates. The sequences correspond to histones H4 and H2A.X. Only one clone from histone H4 and four clones from histone H2A.X

were found indicating the low expression level of this gene, as observed in other dinoflagellate histones (Hackett et al. 2005).

## ***Conclusions***

The presence of plastid-associated genes in the early-divergent heterotrophic dinoflagellate *C. cohnii* suggests that it descends from a plastid-bearing ancestor and may retain an intact, but unrecognized, plastid. This is consistent with an early acquisition of plastids in dinoflagellates, and is compatible with the chromalveolate hypothesis, which proposes a single endosymbiotic event in the common ancestor of alveolates, cryptophytes, haptophytes, and heterokonts, although it does not directly test that hypothesis. *C. cohnii* provides a model for the genomic content of dinoflagellates that have lost plastids, and I present here the first analysis of plastid-related metabolism in a heterotrophic dinoflagellate. The study of other heterotrophic species, particularly those from early-diverging lineages, may reveal a secondary loss of plastids in other organisms. Such data would help determine the timing of plastid acquisition by chl c containing algae.

If, as molecular analyses indicate, the correct phylogenetic position of *C. cohnii* is as an outgroup to photosynthetic dinoflagellates, then *C. cohnii* properties such as the presence of form II rubisco, DOXP isoprenoid biosynthesis, and other plastid-associated genes carry important information concerning overall evolution of dinoflagellate plastids. A prediction from the early acquisition of form II rubisco in dinoflagellates would be that anomalously-pigmented (e.g. fucoxanthin-containing) dinoflagellates may have retained this form of the gene in addition to the plastid-



encoded form I rubisco acquired along with their new plastid. Alternatively, if the phylogenetic placement of *C. cohnii* is within the Gonyaucales as suggested by morphological data, the interpretation would need to be narrowed, but these data still confirm that heterotrophic dinoflagellates would retain genomic evidence of their photosynthetic ancestry.

## Chapter VI – Conclusions and Future Directions

### *Conclusions*

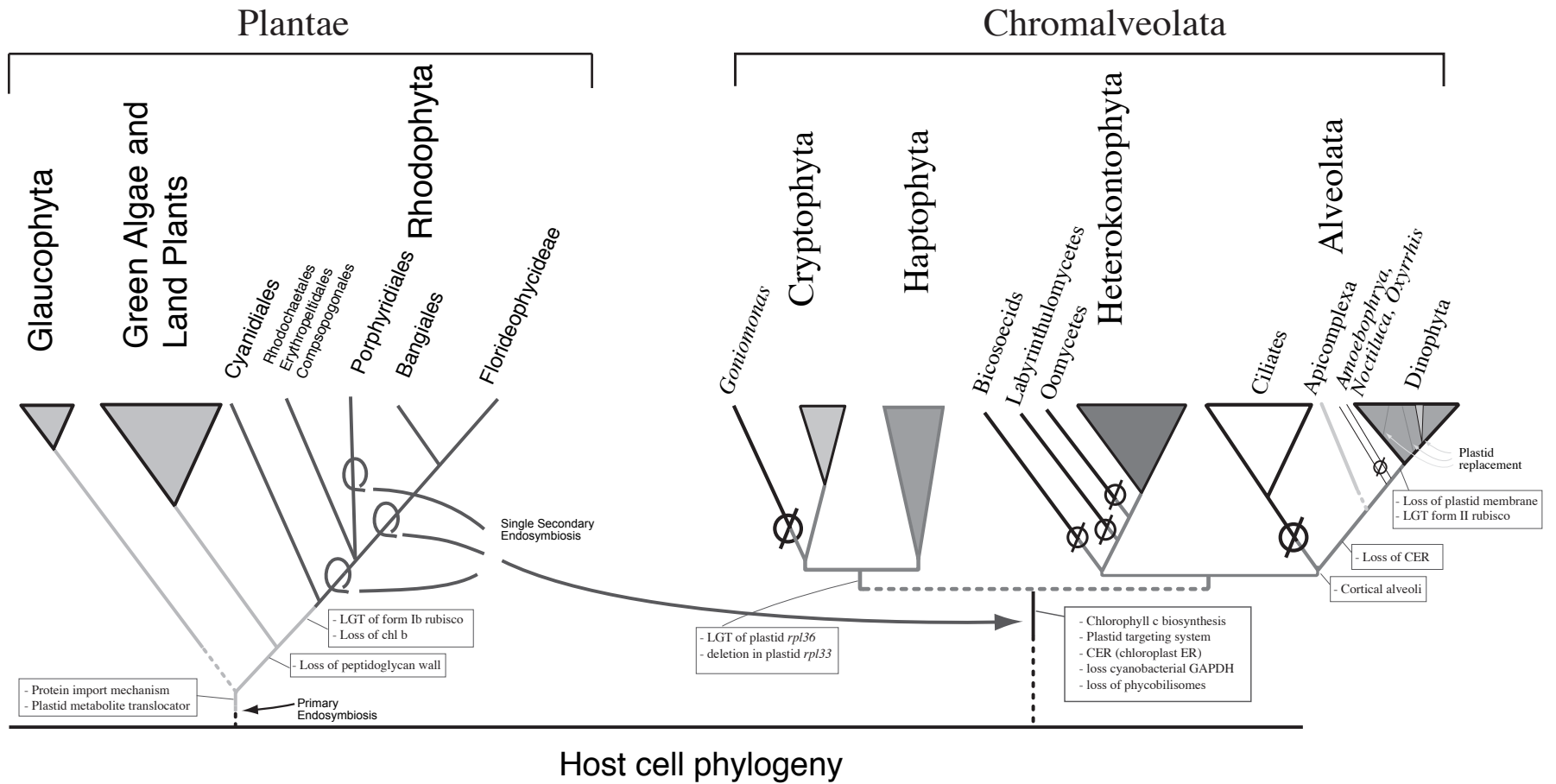
Understanding the evolutionary history of photosynthetic eukaryotes entails the study of phylogenetic relationships among both host cell lineages and their plastids. By comparing phylogenies of eukaryotes and their endosymbionts, an integrated representation of algal evolution would emerge. However, phylogenetic inference of both host cells and plastids has proven challenging, to say the least. The putative rapid radiation of the main eukaryotic lineages may make it difficult to reconstruct deep relationships, and lead to unresolved phylogenies (Knoll 1992; Cavalier-Smith 1999; Baldauf et al. 2000). In addition, molecular datasets often contain spurious, non historical signals arising from compositional bias, heterogeneous rates among lineages, and other patterns, that interfere with phylogenetic analyses and the recovery of the true phylogeny (Ho and Jermiin 2004).

Hypotheses of host cell evolution have been tested using nuclear or mitochondrial molecular data. Nuclear genes are not easily amplified from a wide range of taxa with a single set of primers, and they are often part of gene families, which greatly reduces the number of genes useful for phylogenetic analysis. Phylogenetic studies based on a few concatenated genes were typically able to identify major eukaryotic groups but could not resolve deep phylogenetic relationships (Van de Peer and De Wachter 1997; Baldauf et al. 2000; Baldauf 2003; Cavalier-Smith 2003; Stechmann and Cavalier-Smith 2003).

In contrast, plastid evolution has been studied on the basis of a high number of genes given the feasibility of sequencing complete plastid genomes, although taxon sampling has generally been lacking (Martin et al. 1998; Martin et al. 2002; Bachvaroff et al. 2005 and chapter IV). Some studies, however, have included a denser taxon sampling and only one or a few plastid genes (Fast et al. 2001; Ishida and Green 2002; Yoon et al. 2005). Neither approach has been successful in elucidating all plastid relationships, probably due to heterogeneous evolutionary rates across genes and taxa, compositional bias, low taxon sampling (even in the largest studies), and confounding events of lateral gene transfer. Despite these difficulties, the origin of plastids is today partially understood, while relationships among them remain quite controversial.

A model describing the evolution of several photosynthetic eukaryotes is shown in Figure VI.1. Under this model, a single primary endosymbiotic event gave rise to glaucophytes, green algae (including land plants), and red algae, collectively called Plantae (Cavalier-Smith 1998) or Archaeplastida (Adl et al. 2005). Two main innovations occurred in the common ancestor of Plantae, including the origin of a plastid metabolite translocator necessary for establishment of the organelle (Weber et al. 2006), and the development of a protein import mechanism required after the transfer of most cyanobacterial genes to the host nucleus (Cavalier-Smith 2002; Steiner and Löffelhardt 2002). In congruence with the proposed monophyly of the Plantae, plastids from glaucophytes, red and green algae descend from a common ancestor in cyanobacteria (Delwiche et al. 1995). A sister relationship of red and green algae is strongly supported by molecular data (Burger et al. 1999; Moreira et al.

Figure VI.1. Proposed model of chromalveolate evolution. Dashed lines indicate relationships with weak support. Crossed circles represent putative plastid losses in heterotrophic lineages.



2000; Sanchez-Puerta et al. 2004; Weber et al. 2006 and chapter II); however, the inclusion of glaucophytes within this clade is still debated (Bhattacharya et al. 1995; Nozaki et al. 2003b; Stiller et al. 2003). Recent studies based on molecular data support the inclusion of glaucophytes in Plantae, possibly sister to a clade formed by green and red algae (Moreira et al. 2000; Keeling 2004; Rodriguez-Ezpeleta et al. 2005).

Based on current evidence, including this dissertation work, a model of chromalveolate evolution can be now outlined (Figure VI.1). This model is derived from the so-called “chromalveolate hypothesis” (Cavalier-Smith 1999), and it represents a collection of interrelated hypotheses, each of which can be individually tested. Overall, no data strongly contradicts the chromalveolate hypothesis, and supporting evidence is growing, although relationships among chromalveolate lineages are at best equivocal. Alternative hypotheses (or variations of the current hypothesis) are still viable and should not be lightly disregarded.

The chromalveolate hypothesis proposes a single origin of all four chromalveolate lineages: Cryptophyta, Haptophyta, Heterokontophyta, and Alveolata (Cavalier-Smith 1999). The common ancestor of chromalveolates may have acquired a plastid in a secondary endosymbiotic event by engulfing a unicellular red alga (Cavalier-Smith 1999). Thus, all chromalveolates, including heterotrophic ones, may have evolved from a plastid-bearing ancestor, and may retain a reduced plastid or plastid-derived genes in the nuclear genome. The red algal endosymbiont evolved after the divergence of members of the Cyanidiales, and before the split between Bangiales and Florideophycideae (Yoon et al. 2004 and chapter IV). Better taxon

sampling within the Rhodophyta is needed to establish conclusively the closest living ancestor of chromalveolate plastids.

The model of evolution presented here is congruent with the monophyletic origin of chl *c* plastids found in cryptophytes, haptophytes, heterokonts, and dinoflagellates (Ishida and Green 2002; Yoon et al. 2004; Bachvaroff et al. 2005), although other scenarios may also explain chromalveolate plastid monophyly (Chapter IV). Historically, chromalveolate lineages were grouped together based on plastid features, such as pigmentation or plastid membranes, and later, based on plastid molecular data (Cavalier-Smith 1998; Cavalier-Smith 2002; Palmer 2003). However, it is important to remember that plastid monophyly is congruent with, but not proof of, a monophyletic chromalveolate clade. Today, some molecular data from the host cells partially support the chromalveolate hypothesis.

A sister relationship of heterokonts and alveolates is supported by phylogenetic analyses, and has been recovered using several independent molecular datasets (Van de Peer and De Wachter 1997; Fast et al. 2002; Harper et al. 2005). On the other hand, cryptophytes and haptophytes are more difficult to place, being often associated with different taxa with low support (Bhattacharya et al. 1993; Medlin et al. 1997; Tengs et al. 2000; Stechmann and Cavalier-Smith 2003 and chapter II). Recent phylogenetic analyses based on mitochondrial and nuclear genes showed a weak relationship of haptophytes and cryptophytes (Stechmann and Cavalier-Smith 2003; Harper et al. 2005 and chapter II). In addition, plastids from these two groups appear to be closely related. The plastid gene *rpl36* has been laterally transferred from a bacterium to cryptophyte and haptophyte plastid genomes (J.D. Palmer, pers.

comm.). Furthermore, these two lineages share an eight-amino acid deletion in the plastid gene *rpl33* (chapter III, data not shown). A single lateral gene transfer (*rpl36*) and deletion (in *rpl33*) events may be the simplest explanation for the data, indicating that plastids from cryptophytes and haptophytes are sister taxa. Assuming a sister relationship of cryptophytes and haptophytes, these events may have occurred once in their common ancestor before the divergence of the two lineages. Taken together, the monophyly of alveolates + heterokonts and cryptophytes + haptophytes suggest a chromalveolate clade. Ultrastructural features, such as the shape of mitochondrial cristae, have been used as characters to infer phylogenies (Taylor 1999). All chromalveolates have tubular cristae, except for cryptophytes with flattened cristae, which could be secondarily so (Cavalier-Smith 1998; Taylor 1999). To the best of my knowledge, phylogenetic analyses have not recovered the chromalveolate clade as monophyletic, yet no strong, conflicting relationships have been described either.

The evolutionary model presented here relies partly on the assumption that complex processes are not likely to originate twice in evolution, indicating that certain evolutionary innovations arose once at most (Cavalier-Smith 1999). Some of these innovations include: biosynthesis of chlorophyll *c*, development of a targeting system for complex plastids (surrounded by more than two membranes), and fusion of the outermost plastid membrane with the endoplasmic reticulum to form CER (chloroplast ER). However, the complexity of a process, along with the likelihood of that process to originate only once, are often arbitrarily determined.

Minimizing the occurrence of an evolutionary innovation increases the number of times another evolutionary event is required to happen in order to explain

the data. Under the evolutionary model described here, the number of plastid losses in chromalveolates is large, implying that losing the plastid and its plastid genome (or at least photosynthesis) may occur readily in evolution. Interestingly, no heterotrophic organism unquestionably derived from a plastid-bearing ancestor has been shown to have completely lost its plastid, leaving no traces. The apicomplexan parasite *Cryptosporidium parvum* retained at least a few plastid-derived genes in its nuclear genome (Huang et al. 2004), and the heterotrophic dinoflagellate *Cryptothecodinium cohnii* contain plastid-derived genes in the its nucleus and probably a reduced plastid (Chapter V). Parasitic plants, including hemi and holoparasitic species, still retain a leucoplast with a plastid genome (Wolfe et al. 1992; Bungard 2004; de Koning and Keeling 2004). Some heterotrophic chromalveolates were found to contain a remnant plastid when studied in detail (Sekiguchi et al. 2002; Foth and McFadden 2003). Therefore, the presence of reduced plastids and/or plastid-derived genes in the nuclear genome of basal heterotrophic cryptophytes, heterokonts, and alveolates is expected, or complete loss of plastids and evidence of the past presence of one in those lineages will demand further explanations.

An even more complex evolutionary history is required to explain plastid diversity in Alveolata. Apicomplexan plastids are surrounded by four membranes not including a CER (Wilson 2002; Foth and McFadden 2003). The origin of apicoplasts is contentious (Funes et al. 2004), with stronger support for red algal ancestry (Fast et al. 2001), although the relationship with peridinin-containing plastids from dinoflagellates is more controversial. Most photosynthetic dinoflagellates possess plastids pigmented with peridinin and surrounded by three membranes (Delwiche



1999; Bhattacharya et al. 2003). In addition, peridinin dinoflagellates contain a nuclear-encoded form II rubisco laterally transferred from proteobacteria (Morse et al. 1995). Several dinoflagellates replaced the peridinin-containing plastid with other plastid types, including some acquired from green algae, haptophytes, and diatoms in independent secondary or tertiary endosymbiotic events (Dodge 1975; Chesnick et al. 1997; Schnepf and Elbrächter 1999; Tengs et al. 2001). Abnormally pigmented dinoflagellates contain plastid-encoded form I rubisco derived from the latest endosymbiont, although it is possible they have also retained form II rubisco in their nuclear genome. It remains unknown whether these “new plastids” represent true organelles that evolved protein-import mechanisms, or are obligate endosymbionts (Cavalier-Smith 1999). The most prominent group of anomalously pigmented dinoflagellates are those containing haptophyte-derived tertiary plastids. Preliminary studies of protein trafficking in haptophyte-containing dinoflagellates revealed a similar protein targeting mechanism as peridinin-containing dinoflagellates (Patron et al. 2005; Patron et al. 2006). Further work needs to be done on other anomalously pigmented dinoflagellates to learn whether a protein transport system has been established and if so, whether a novel system was developed, or the same protein import mechanism is maintained.

### ***Future Directions***

The importance of understanding evolutionary relationships has been previously stressed (Yates et al. 2004). Most relevant implications include the development of novel treatments for a particular disease by learning the evolutionary history of the causative agent. The most notorious example are *Plasmodium*

*falciparum*, and other apicomplexan parasites that are responsible for malaria, toxoplasmosis, coccidiosis, among others. Today, most drug targets to fight malaria include pathways localized to (or derived from) the endosymbiont (apicoplast) present in apicomplexans. Of particular importance would be to understand the evolution of parasitic oomycetes, e.g. *Phytophthora infestans*, causative agent of the potato late blight. Finding evidence of an earlier endosymbiosis, may uncover relevant metabolic pathways that could be useful to manage the parasite.

Studying heterotrophic chromalveolates will also help us understand the effects of plastid loss or reduction, including retention of plastid-associated genes in the nuclear genome and maintenance of plastid-derived metabolism. An interesting aspect to study would be the re-localization of plastid-specific pathways in secondarily heterotrophic organisms that lack these organelles. As I mentioned before, most genes from the endosymbiont are transferred to the host nuclear genome and proteins encoded by those genes are targeted back to the plastid (Martin and Herrmann 1998; Martin et al. 1998). However, some plastid-derived genes are relocated and targeted to different cellular compartments, other than the plastid. Only a few examples of plastid proteins functioning in other compartments are known so far, and a number of others have been suggested using target prediction programs (Brinkmann and Martin 1996; Martin and Herrmann 1998). To fully understand the complexity of the endosymbiotic event and the chimeric origin of photosynthetic eukaryotes, an extensive study of the localization of all plastid-derived genes that reside in the host nucleus should be conducted, including detailed phylogenetic analyses and identification of putative targeting signals complemented with bench

experiments. Last but not least, understanding chromalveolate evolution would be useful to interpret relative fitness and success of these photosynthetic eukaryotes living in different environmental conditions, and how they influence the overall marine ecosystem.

## References

- Adl, S., Simpson, A., Farmer, M. et al. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Euk. Microbiol.* 52:399-451.
- Ahmad, I. & Rao, D. N. 1996. Chemistry and biology of DNA methyltransferases. *Crit. Rev. Biochem. Mol.* 31:361-380.
- Ali, A. B., Baere, R. D., Auwera, G. V. d., Wachter, R. D. & Van de Peer, Y. 2001. Phylogenetic relationships among algae based on complete large-subunit rRNA sequences. *International Journal of Systematics and Evolutionary Microbiology* 51:737-749.
- Allen, J. & Raven, J. 1996. Free-radical-induced mutation vs. redox regulation: costs and benefits of genes in organelles. *J. Mol. Evol.* 42:482-492.
- Andersen, R. A. 1991. The cytoskeleton of chromophyte algae. *Protoplasma* 164:143-159.
- Andersen, R. A., Morton, S. L. & Sexton, J. P. 1997. Provasoli-Guillard National Center for Culture of Marine Phytoplankton 1997 list of strains. *J. Phycol.* 33:1-75.
- Andersen, R. A. 2004. Biology and systematics of heterokont and haptophyte algae. *Am. J. Bot.* 91:1508-1522.
- Andersson, J. O. & Roger, A. J. 2002. A cyanobacterial gene in nonphotosynthetic protists--an early chloroplast acquisition in eukaryotes? *Curr. Biol.* 12:115-119.

- Ane, C., Burleigh, J., McMahon, M. & Sanderson, M. 2005. Covarion structure in plastid genome evolution: a new statistical test. *Mol. Biol. Evol.* 22:914-924.
- Armbrust, E. V., Berges, J., Bowler, C. et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306:79-86.
- Bachvaroff, T. R., Concepcion, G. T., Rogers, C. R. & Delwiche, C. F. 2004. Dinoflagellate EST data indicate massive transfer of chloroplast genes to the nucleus. *Protist* 155:65-78.
- Bachvaroff, T. R., Sanchez-Puerta, M. V. & Delwiche, C. F. 2005. Chlorophyll *c* containing plastid relationships based on analyses of a multi-gene dataset with all four chromalveolate lineages. *Mol. Biol. Evol.* 22:1-11.
- Bachvaroff, T. R., Sanchez-Puerta, M. V. & Delwiche, C. F. 2006. Rate variation as a function of gene origin in plastid-derived genes of peridinin-containing dinoflagellates. *J. Mol. Evol.* 62:42-52.
- Balch, W. M., Holligan, P. M. & Kilpatrick, K. A. 1992. Calcification, photosynthesis and growth of the bloom-forming coccolithophore, *Emiliania huxleyi*. *Cont. Shelf Res.* 12:1353-1374.
- Baldauf, S. F. 2003. The deep roots of eukaryotes. *Science* 300:1703-1706.
- Baldauf, S. F., Bhattacharya, D., Cockrill, J., Hugenholtz, P., Pawlowski, J. & Simpson, A. 2004. The tree of life. In Cracraft, J., and Donoghue, M. Eds. *Assembling the Tree of Life*. Oxford University Press, New York, pp. 43-75.

- Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. & Doolittle, W. F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290:972-976.
- Barbrook, A. & Howe, C. 2000. Minicircular plastid DNA in the dinoflagellate *Amphidinium operculatum*. *Mol. Gen. Genet.* 263:152-158.
- Bendtsen, J., Nielsen, H., von Heijne, G. & Brunak, S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340:783-795.
- Bhattacharya, D., Stickel, S. K. & Sogin, M. L. 1993. Isolation and molecular phylogenetic analysis of actin-coding regions from *Emiliana huxleyi*, a prymnesiophyte alga, by reverse transcriptase and PCR methods. *Mol. Biol. Evol.* 10:689-703.
- Bhattacharya, D., Helmchen, T., Bibeau, C. & Melkonian, M. 1995. Comparisons of nuclear-encoded small-subunit ribosomal RNAs reveal the evolutionary position of the Glaucocystophyta. *Mol. Biol. Evol.* 12:415-420.
- Bhattacharya, D., Yoon, H. S. & Hackett, J. D. 2003. Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *BioEssays* 26:50-60.
- Bijornland, T. & Liaaen-Jensen, S. 1989. Distribution patterns of carotenoid in relation to chromophyte phylogeny and systematics. In Green, J. C., Leadbeater, B. S. C., and Diver, W. Eds. *The Chromophyte algae: problems and perspectives*. Clarendon Press, Oxford, pp. 37-60.
- Bock, H., Brennicke, A. & Schuster, W. 1994. *Rps3* and *rpl16* genes do not overlap in *Oenothera* mitochondria: GTG as a potential translation initiation codon in plant mitochondria? *Plant Mol. Biol.* 14:811-818.

- Borza, T., Popescu, C. & Lee, R. 2005. Multiple metabolic roles for the nonphotosynthetic plastid of the green alga *Prototheca wickerhamii*. *Eukaryot. Cell* 4:253-261.
- Bourrelly, P. 1957. Recherches sur les Chrysophycées: morphologie, phylogenie, systematique. *Rev. Algol. Mém. Hors-Séri* 1:1-412.
- Brasier, M., Green, O., Jephcoat, A., Kleppe, A., van Kranendonk, M., Lindsay, J., Steele, A. & Grassineau, N. 2002. Questioning the evidence for Earth's oldest fossils. *Nature* 416:76-81.
- Brinkmann, H. & Martin, W. 1996. Higher plant chloroplast and cytosolic 3-phosphoglycerate kinases: a case of endosymbiotic gene replacement. *Plant Mol. Biol.* 30:65-75.
- Brown, C. W. & Yoder, J. A. 1994. Coccolithophorid blooms in the global oceans. *J. Geophys. Res.* 99:7467-7482.
- Brown, J. R. 2003. Ancient horizontal gene transfer. *Nature Reviews* 4:121-132.
- Buchanan, B., Gruissem, W. & Jones, R. 2000. Biochemistry and molecular biology of plants. American Society of Plant Physiologists, Rockville, MD, 1363 pp.
- Bungard, R. 2004. Photosynthetic evolution in parasitic plants: insight from the chloroplast genome. *BioEssays* 36:235-247.
- Burger, G., Plante, I., Lonergan, K. M. & Gray, M. W. 1995. The mitochondrial genome of the amoeboid protozoon, *Acanthamoeba castellanii*. Complete sequence, gene content and genome organization. *J. Mol. Biol.* 239:476-499.

- Burger, G., Saint-Louis, D., Gray, M. W. & Lang, B. F. 1999. Complete sequence of the mitochondrial DNA of the red alga *Porphyra purpurea*: cyanobacterial introns and shared ancestry of red and green algae. *Plant Cell* 11:1675-1694.
- Burger, G., Zhu, Y., Littlejohn, T. G., Greenwood, S. J., Schnare, M. N., Lang, B. F. & Gray, M. W. 2000. Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aurelia* mitochondrial DNA. *J. Mol. Biol.* 297:365-380.
- Burger, G., Lang, B. F., Braun, H. P. & Marx, S. 2003. The enigmatic mitochondrial ORF *yfm39* codes for ATP synthase chain b. *Nucleic Acids Res.* 31:2353-2360.
- Cai, X., Fuller, L., McDougald, L. & Zhu, G. 2003. Apicoplast genome of the coccidian *Eimeria tenella*. *Gene* 321:39-46.
- Capowski, E. E., Wells, J. M., Harrison, G. S. & Karrer, K. M. 1989. Molecular analysis of N6-Methyladenine patterns in *Tetrahymena thermophila* nuclear DNA. *Mol. Cell. Biol.* 9:2598-2605.
- Cavalier-Smith, T. 1981. Eukaryote kingdoms: seven or nine? *BioSystems* 14:461-481.
- Cavalier-Smith, T. 1986. The kingdom Chromista: origin and systematics. *Progress Phycological Res.* 4:310-347.
- Cavalier-Smith, T. 1989. The kingdom Chromista. In Green, J., Leadbeater, B. S. C., and Diver, W. Eds. *The Chromophyte Algae: problems and perspectives*. Clarendon Press, Oxford, pp. 381-407.



- Cavalier-Smith, T. 1991. Cell diversification in heterotrophic flagellates. *In* Patterson, D., and Larsen, J. Eds. *The biology of free living heterotrophic flagellates*. Oxford University Press, New York, pp. 113-131.
- Cavalier-Smith, T., Allsopp, M. & Chao, E. 1994. Chimeric conundra: are nucleomorphs and chromists monophyletic or polyphyletic? *Proc. Natl. Acad. Sci. USA* 91:11368-11372.
- Cavalier-Smith, T. & Chao, E. 1996. 18S rRNA sequence of *Heterosigma carterae* (Raphidophyceae), and the phylogeny of heterokont algae (Ochrophyta). *Phycologia* 35:500-510.
- Cavalier-Smith, T. 1998. A revised six-kingdom system of life. *Biol. Rev.* 73:203-266.
- Cavalier-Smith, T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J. Eukaryot. Microbiol.* 46:347-366.
- Cavalier-Smith, T. 2000. Membrane heredity and early chloroplast evolution. *Trends Plant Sci.* 5:174-182.
- Cavalier-Smith, T. 2002. Genomic reduction and evolution of novel membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). *Phil. Trans. R. Soc. Lond.* 358:109-134.
- Cavalier-Smith, T. 2003. Protist phylogeny and the high level classification of Protozoa. *Europ. J. Protistol.* 39:338-348.
- Cavalier-Smith, T. 2004. Only six kingdoms of life. *Proc. R. Soc. Lond. B* 271:1251-1262.

- Chapdelaine, Y. & Bonen, L. 1991. The wheat mitochondrial gene for subunit I of the NADH dehydrogenase complex: A *trans*-splicing model for this gene-in-pieces. *Cell* 65:465-472.
- Chesnick, J. M. & Cattolico, R. A. 1993. Isolation of DNA from eukaryotic algae. *Methods Enzymol.* 224:168-176.
- Chesnick, J. M., Kooistra, W. H. C. F., Wellbrock, U. & Medlin, L. K. 1997. Ribosomal RNA analysis indicates a benthic pennate diatom ancestry for the endosymbionts of the dinoflagellates *Peridinium foliaceum* and *Peridinium balticum* (Pyrrhophyta). *J. Euk. Microbiol.* 44:314-320.
- Christensen, T. 1962. II. Systematisk Botanik, 2. Alger. In Bocher, T., Lange, M., and Sorensen, T. Eds. *Botanik*, Munksgaard, Copenhagen, pp. 1-178.
- Christensen, T. 1989. The Chromophyta, past and present. In Green, J. C., Leadbeater, B. S. C., and Diver, W. Eds. *The Chromophyte Algae: Problems and Perspectives*. Clarendon Press, Oxford, pp. 1-12.
- Chu, K., Qi, J., Yu, Z.-G. & Anh, V. 2004. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Mol. Biol. Evol.* 21:200-206.
- Conklin, P. L., Wilson, R. K. & Hanson, M. R. 1991. Multiple *trans*-splicing events are required to produce a mature *nad1* transcript in a plant mitochondrion. *Gene. Dev.* 5:1407-1415.
- Cornah, J., Terry, M. & Smith, A. 2003. Green or red: what stops the traffic in the tetrapyrrole pathway? *Trends Plant Sci.* 8:224-230.

- Crick, F. H. C. 1966. Codon-anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* 19:548-555.
- Cunillera, N., Boronat, A. & Ferrer, A. 2000. Spatial and temporal patterns of GUS expression directed by 5' regions of the *Arabidopsis thaliana* farnesyl diphosphate synthase genes *FPS1* and *FPS2*. *Plant Mol. Biol.* 44:747-758.
- Cunningham, C. & Gantt, E. 1998. Genes and enzymes of carotenoid biosynthesis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 49:557-583.
- Cunningham, C., Lafond, T. & Gantt, E. 2000. Evidence of a role for *LytB* in the nonmevalonate pathway of isoprenoid biosynthesis. *J. Bacteriology* 182:5841-5848.
- Daugbjerg, N. & Andersen, R. A. 1997. Phylogenetic analyses of the *rbcL* sequences from haptophytes and heterokont algae suggest their chloroplasts are unrelated. *Mol. Biol. Evol.* 14:1242-1251.
- de Koning, A. & Keeling, P. J. 2004. Nucleus-encoded genes for plastid-targeted proteins in *Helicosporidium*: functional diversity of a cryptic plastid in a parasitic alga. *Eukaryot. Cell* 3:1198-1205.
- Delwiche, C. F., Kuhsel, M. & Palmer, J. D. 1995. Phylogenetic analysis of *tufA* sequences indicates a cyanobacterial origin of all plastids. *Mol. Phylogenet. Evol.* 4:110-1288.
- Delwiche, C. F. & Palmer, J. D. 1996. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. *Mol. Biol. Evol.* 13:873-882.
- Delwiche, C. F. & Palmer, J. D. 1997. The origin of plastids and their spread via secondary symbiosis. *Plant Syst. Evol.* 11:S53-S86.

- Delwiche, C. F. 1999. Tracing the thread of plastid diversity through the tapestry of life. *Am. Nat.* 154:S164-S177.
- Disch, A., Schwender, J., Muller, C., Lichtenthaler, H. & Rohmer, M. 1998. Distribution of the mevalonate and glyceraldehyde phosphate/pyruvate pathways for isoprenoid biosynthesis in unicellular algae and the cyanobacterium *Synechocystis* PCC 6714. *Biochem. J.* 333:381-388.
- Dodge, J. D. 1975. A survey of chloroplast ultrastructure in the dinophyceae. *Phycologia* 14:253-263.
- Douglas, S. E. & Penny, S. L. 1999. The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved syntenic groups confirm its common ancestry with red algae. *J. Mol. Evol.* 48:236-244.
- Durnford, D., Deane, J., Tan, S., McFadden, G. I., Gantt, E. & Green, B. 1999. A phylogenetic assessment of the eukaryotic light-harvesting antenna protein, with implications for plastid evolution. *J. Mol. Evol.* 48:59-68.
- Edwardsen, B., Eikrem, W., Green, J. C., Andersen, R. A., Moon-van der Staay, S. Y. & Medlin, L. K. 2000. Phylogenetic reconstruction of the Haptophyta inferred from 18S ribosomal DNA sequences and available morphological data. *Phycologia* 39:19-35.
- Edwards, A. 1972. Likelihood. Cambridge University Press, Cambridge, 235 pp.
- Emanuelsson, O., Nielsen, H. & von Heijne, G. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Prot. Sci.* 8:978-984.

- Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300:1005-1016.
- Emanuelsson, O. & Heijne, G. v. 2001. Prediction of organellar targeting signals. *Biochimica et Biophysica Acta* 1541:114-119.
- Eschbach, S., Hofmann, C., Maier, U.-G., Sitte, P. & Hansmann, P. 1991. A eukaryotic genome of 660 kb: electrophoretic karyotype of nucleomorph and cell nucleus of the cryptomonad alga, *Pyrenomonas salina*. *Nucleic Acids Res.* 19:1779-1781.
- Falkowski, P., Scholes, R. J., Boyle, E. et al. 2000. The global carbon cycle: a test of our knowledge of earth as a system. *Science* 290:291-296.
- Falkowski, P., Katz, M., Knoll, A., Quigg, A., Raven, J., Schofield, O. & Taylor, F. 2004. The evolution of modern eukaryotic phytoplankton. *Science* 305:354-360.
- Falkowski, P. 2006. Tracing oxygen's imprint on Earth's metabolic evolution. *Science* 311:1724-1725.
- Fast, N. M., Kissinger, J. C., Roos, D. S. & Keeling, P. J. 2001. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol. Biol. Evol.* 18:418-426.
- Fast, N. M., Xue, L., Bingham, S. & Keeling, P. J. 2002. Re-examining alveolate evolution using multiple protein molecular phylogenies. *J. Euk. Microbiol.* 49:30-37.

- Fedoreyeva, L. I. & Vanyushin, B. F. 2002. N6-Adenine DNA-methyltransferase in wheat seedlings. *FEBS Letters* 514:305-308.
- Fensome, R., Saldarriaga, J. F. & Taylor, F. 1999. Dinoflagellate phylogeny revisited: reconciling morphological and molecular based phylogenies. *Grana* 38:66-80.
- Foth, B. J. & McFadden, G. I. 2003. The apicoplast: a plastid in *Plasmodium falciparum* and other apicomplexan parasites. *International Review of Cytology* 224:57-110.
- Fujiwara, S., Tsuzuki, M., Kawachi, M., Minaka, N. & Inouye, I. 2001. Molecular phylogeny of the Haptophyta based on the *rbcL* gene and sequence variation in the spacer region of the Rubisco operon. *J. Phycol.* 37:121-129.
- Funes, S., Reyes-Prieto, A., Perez-Martinez, X. & Gonzalez-Halphen, D. 2004. On the evolutionary origins of apicoplasts: revisiting the rhodophyte vs. chlorophyte controversy. *Microbes and Infection* 6:305-311.
- Gardner, M., Williamson, D. & Wilson, R. J. M. 1. 1991. A circular DNA in malaria parasites encodes an RNA polymerase like that of prokaryotes and chloroplasts. *Mol. Biochem. Parasitol.* 44:115-123.
- Gibbs, S. 1962. Nuclear envelope-chloroplast relationships in algae. *Journal of Cell Biology* 14:433-444.
- Gibbs, S. 1970. The comparative ultrastructure of the algal chloroplast. *Ann. N.Y. Acad. Sci.* 175:454-473.
- Gibbs, S. 1978. The chloroplasts of *Euglena* may have evolved from symbiotic green algae. *Can. J. Bot.* 56:2883-2889.

- Gibbs, S. 1981a. The chloroplast of some algal groups may have evolved from endosymbiotic eukaryotic algae. *Ann. N.Y. Acad. Sci.* 361:193-208.
- Gibbs, S. 1981b. The chloroplast endoplasmic reticulum: structure, function, and evolutionary significance. *Int. Rev. Cytol.* 72:49-99.
- Gilson, P. & McFadden, G. I. 1996. The miniaturized nuclear genome of a eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns. *Proc. Natl Acad. Sci. USA* 93:7737-7742.
- Gilson, P. 2001. Nucleomorph genomes: much ado about practically nothing. *Genome Biology* 2:1022.
- Glöckner, G., Rosenthal, A. & Valentin, K. 2000. The structure and gene repertoire of an ancient red algal plastid genome. *J. Mol. Evol.* 51:382-390.
- Gockel, G. & Hachtel, W. 2000. Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*. *Protist* 151:347-351.
- Gogarten, J., Kibak, H., Dittrich, P. et al. 1989. Evolution of the vacuolar H<sup>+</sup>-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. USA* 86:6661-6665.
- Goremykin, V., Hirsch-Ernst, K., Wolfl, S. & Hellwig, F. 2004. The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. *Mol. Biol. Evol.* 21:1445-1454.
- Graham, L. & Wilcox, L. 2000. *Algae*. Prentice-Hall, Inc., Upper Saddle River, NJ, 640 pp.

- Gray, M. W., Lang, B. F., Cedergren, R. et al. 1998. Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res.* 26:865-878.
- Gray, M. W., Burger, G. & Lang, B. F. 1999. Mitochondrial evolution. *Science* 283:1476-1481.
- Green, B. R. & Jordan, R. 1994. Systematic history and taxonomy. In Green, J. C., and Leadbeater, B. S. C. Eds. *The Haptophyte Algae*. Clarendon Press, Oxford, pp. 1-21.
- Green, J. C. & Harris, R. P. 1996. EHUX (*Emiliania huxleyi*). *J. Marine Syst.* 9:1-136.
- Grzebyk, D., Schofield, O., Vetriani, C. & Falkowski, P. G. 2003. The mesozoic radiation of eukaryotic algae: the portable plastid hypothesis. *J. Phycol.* 39:1-10.
- Guillot, M. & Gibbs, S. 1980a. The cryptomonad nucleomorph: its ultrastructure and evolutionary significance. *J Phycol* 16:558-568.
- Guillot, M. & Gibbs, S. 1980b. Evidence that the chloroplast and nucleomorph of cryptomonads are remnants of a eukaryotic symbiont. *J Cell Biol* 87:186.
- Guindon, S. & Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52:696-704.
- Gunderson, J., Goss, S. & Coats, D. W. 1999. The phylogenetic position of *Amoebophrya* sp infecting *Gymnodinium sanguineum*. *J. Euk. Microbiol.* 46:194-197.



- Hackett, J. D., Yoon, H. S., Soares, M. B., Bonaldo, M. F., Casavant, T. L., Scheetz, T. E., Nosenko, T. & Bhattacharya, D. 2004. Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. *Curr. Biol.* 14:213-218.
- Hackett, J. D., Scheetz, T. E., Yoon, H. S., Soares, M. B., Bonaldo, M. F., Casavant, T. L. & Bhattacharya, D. 2005. Insight into a dinoflagellate genome through expressed sequence tag analysis. *BMC Genomics* 6:80-93.
- Hagopian, J., Reis, M., Kitajima, J., Bhattacharya, D. & Oliveira, M. 2004. Comparative analysis of the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. *liui* provides insights into the evolution of rhodoplasts and their relationship to other plastids. *J. Mol. Evol.* 59:464-477.
- Hallick, R., Hong, L., Drager, R., Favreau, M., Monfort, A., Orsat, B., Spielmann, A. & Stutz, E. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res.* 21:3537-3544.
- Hanson, M. R. & Folker, T. S. 1992. Structure and function of the higher plant mitochondrial genome. *Int. Rev. Cytol.* 141:129-172.
- Harper, J. T. & Keeling, P. J. 2003. Nucleus-encoded, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase (GAPDH) indicates a single origin for chromalveolate plastids. *Mol. Biol. Evol.* 20:1730-1735.
- Harper, J. T., Waanders, E. & Keeling, P. J. 2005. On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. *Int. J. Syst. Evol. Microbiol.* 55:487-496.

- Hayashi-Ishimaru, Y., Ehara, M., Inagaki, Y. & Ohama, T. 1997. A deviant mitochondrial genetic code in prymnesiophytes (yellow-algae): UGA codon for tryptophan. *Curr. Genet.* 32:296-299.
- Heithhoff, D. M., Soinsheimer, R. L., Low, D. A. & Mahan, M. J. 1999. An essential role for DNA adenine methylation in bacterial virulence. *Science* 284:967-970.
- Hiller, R. 2001. Empty minicircles and *petB/atpA* and *psbD/psbE* genes in tandem in *Amphidinium carterae* plastid DNA. *FEBS Letters* 505:449-452.
- Ho, S. & Jermiin, L. 2004. Tracing the decay of the historical signal in the biological sequence data. *Syst. Biol.* 53:623-637.
- Hong, Y. G., Dover, S. L., Cole, T. E., Brasier, C. M. & Buck, K. W. 1999. Multiple mitochondrial viruses in an isolate of the Dutch elm disease fungus *Ophiostoma novo-ulmi*. *Virology* 258:118-127.
- Huang, J., Mullapudi, N., Lancto, C., Scott, M., Abrahamsen, M. & Kissinger, J. C. 2004. Phylogenomic evidence supports ast endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome Biology* 5:R88.
- Huelsenberg, J. P. & Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754-755.
- Inagaki, Y., Ehara, M., Watanabe, K. I., Hayashi-Ishimaru, Y. & Ohama, T. 1998. Directionally evolving genetic code: the UGA codon from stop to tryptophan in mitochondria. *J. Mol. Evol.* 47:378-384.

- Inagaki, Y., Simpson, A., Dacks, J. & Roger, A. J. 2004a. Phylogenetic artifacts can be caused by leucine, serine, and arginine codon usage heterogeneity: dinoflagellate plastid origins as a case study. *Syst. Biol.* 53:582-593.
- Inagaki, Y., Susko, E., Fast, N. M. & Roger, A. J. 2004b. Covarion shifts cause a long-branch attraction artifact that unites Microsporidia and Archaeobacteria in EF-1alpha phylogenies. *Mol. Biol. Evol.* 21:1340-1349.
- Inouye, I. & Kawachi, M. 1994. The haptonema. In Green, J., and Leadbeater, B. S. C. Eds. *The haptophyte algae*. Clarendon Press, Oxford, pp. 73-90.
- Ishida, K. I. & Green, B. R. 2002. Second- and third-hand chloroplasts in dinoflagellates: Phylogeny of oxygen-evolving enhancer1 (*PsbO*) protein reveals replacement of a nuclear-encoded plastid gene by that of a haptophyte tertiary endosymbiont. *Proc. Natl. Acad. Sci. USA* 99:9294-9299.
- Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. 1989. Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* 86:9355-9359.
- Javornicky, P. 1962. Two scarcely known genera of the class Dinophyceae: *Bernardinium* Chodat and *Crypthecodinium* Biecheler. *Preslia* 34:98-113.
- Jermiin, L., Ho, S., Ababneh, F., Robinson, J. & Larkum, A. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst. Biol.* 53:638-643.
- Jones, D., Taylor, W. & Thornton, J. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8:275-282.

- Kaneko, A., Sato, N., Kotani, H. et al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3:109-136.
- Karol, K., McCourt, R., Cimino, M. & Delwiche, C. F. 2001. The closest living relatives of land plants. *Science* 294:2351-2353.
- Karpov, S., Sogin, M. L. & Silberman, J. 2001. Rootlet homology, taxonomy, and phylogeny of bicoseocids based on 18S rRNA gene sequences. *Protistology* 2:34-47.
- Kay, P. H., Pereira, E. & Marlow, S. A. 1994. Evidence for adenine methylation within the mouse myogenic gene MYP-D1. *Gene* 151:89-95.
- Keeling, P. J., Palmer, J. D., Donald, R., Roos, D. S., Waller, R. & McFadden, G. I. 1999. Shikimate pathway in apicomplexan parasites. *Nature* 397:219-220.
- Keeling, P. J. 2004. Diversity and evolutionary history of plastids and their hosts. *Am. J. Bot.* 91:1481-1493.
- Kilian, O. & Kroth, P. 2005. Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids. *Plant J.* 41:175-183.
- Kim, K.-J. & Lee, H.-L. 2004. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res.* 11:247-261.
- Knoll, A. 1992. The early evolution of eukaryotes: a geological perspective. *Science* 256:622-627.

- Köhler, S., Delwiche, C. F., Denny, P. W., Tilney, L. G., Webster, P., Wilson, R., Palmer, J. D. & Roos, D. S. 1997. A plastid of probable green algal origin in apicomplexan parasites. *Science* 275:1485-1489.
- Kolodner, R. & Twarei, K. 1979. Inverted repeats in chloroplast DNA from higher plants. *Proc. Natl. Acad. Sci. USA* 76:41-45.
- Kowallik, K. V., Stoebe, B., Schaffran, I., Kroth-Pancic, P. & Freier, U. 1995. The chloroplast genome of a chlorophyll a+c containing alga, *Odontella sinensis*. *Plant Mol. Biol. Rep.* 13:336-342.
- Kroth, P. 2002. Protein transport into secondary plastids and the evolution of primary and secondary plastids. *International Review of Cytology* 221:191-255.
- Kubai, D. & Ris, H. 1969. Division in the dinoflagellate *Gyrodinium cohnii* (Schiller). *The Journal of Cell Biology* 40:508-528.
- Laatsch, T., Zauner, S., Stoebe-Maier, B., Kowallik, K. V. & Maier, U.-G. 2004. Plastid-derived single gene minicircles of the dinoflagellate *Ceratium horridum* are localized in the nucleus. *Mol. Biol. Evol.* 21:1318-1322.
- Lake, J. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralineal distances. *Proc. Natl. Acad. Sci. USA* 91:1455-1459.
- Lang, B. F., Burger, G., O'Kelly, C. J., Cedergren, R., Golding, G. B., Lemieux, C., Sankoff, D., Turmel, M. & Gray, M. W. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* 387:493-496.
- Lange, B., Rujan, T., Martin, W. & Croteau, R. 2000. Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proc. Natl. Acad. Sci. USA* 97:13172-13177.

- Leander, B. & Keeling, P. J. 2004. Early evolutionary history of dinoflagellates and apicomplexans (Alveolata) as inferred from *hsp90* and actin phylogenies. *J. Phycol.* 40:341-350.
- Leblanc, C., Boyen, C., Richard, O., Bonnard, G., Grienemberger, J. M. & Kloareg, B. 1995. Complete sequence of the mitochondrial DNA of the rhodophyte *Chondrus crispus* (Gigartinales). Gene content and genome organization. *J. Mol. Biol.* 250:484-495.
- Lemieux, C., Otis, C. & Turmel, M. 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature* 403:649-652.
- Li, H.-m. & Chen, L.-J. 1996. Protein targeting and integration signal for the chloroplastic outer envelope membrane. *Plant Cell* 8:2117-2126.
- Lichtenthaler, H., Schwender, J., Disch, A. & Rohmer, M. 1997. Biosynthesis of isoprenoids in higher plant chloroplasts proceeds via a mevalonate-independent pathway. *FEBS Letters* 400:271-274.
- Lidie, K., Ryan, J., Barbier, M. & Van Dolah, F. 2005. Gene expression in Florida red tide dinoflagellate *Karenia brevis*: analysis of an expressed sequence tag library and development of DNA microarray. *Mar. Biotechnol.* 7:481-493.
- Lippmeier, J., Brown, A. M. & Apt, K. E. 2002. Isolation of algal genes by functional complementation of yeast. *J. Phycol.* 38:529-533.
- Litaker, R., Tester, P., Colorni, A., Levy, M. & Noga, E. 1999. The phylogenetic relationship of *Pfiesteria piscicida*, cryptoperidiniopsoid sp. *Amyloodinium*

- ocellatum* and a *Pfiesteria*-like dinoflagellate to other dinoflagellates and apicomplexans. *J. Phycol.* 35:1379-1389.
- Lockhart, P., Steel, M., Barbrook, A., Huson, D., Charleston, M. & Howe, C. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol. Biol. Evol.* 15:1183-1188.
- Lockhart, P., Howe, C., Barbrook, A., Larkum, A. & Penny, D. 1999. Spectral analysis, systematic bias, and the evolution of chloroplasts. *Mol. Biol. Evol.* 16:573-576.
- Lopez-Garcia, P., Rodriguez-Valera, F., Pedros-Alios, C. & Moreira, D. 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* 409:603-607.
- Lowe, T. & Eddy, S. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955-964.
- Maddison, W. & Maddison, P. 2000. MacClade version 4: analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, MA, pp.
- Maier, R., Neckermann, K., Igloi, G. & Kossel, H. 1995. Complete sequence of the maize chloroplast genome-gene content, hotspots of divergence and fine-tuning of genetic information by transcript editing. *J. Mol. Biol.* 251:614-628.
- Maier, U.-G., Douglas, S. & Cavalier-Smith, T. 2000. The nucleomorph genomes of Cryptophytes and Chlorarachniophytes. *Protist* 151:103-109.
- Malik, H. & Henikoff, S. 2003. Phylogenomics of the nucleosome. *Nat. Struct. Biol.* 10:882-891.

- Malin, G., Turner, S. & Liss, P. 1992. Sulfur: the plankton/climate connection. *J. Phycol.* 28:590-597.
- Malin, G. & Kirst, G. O. 1997. Algal production of dimethyl sulfide and its atmospheric role. *J. Phycol.* 33:889-896.
- Malone, T., Blumenthal, R. & Cheng, X. 1995. Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyl-transferases, and suggests a catalytic mechanism for these enzymes. *J. Mol. Biol.* 253:618-632.
- Marchant, H. & Thomsen, H. A. 1994. Haptophytes in polar waters. In Green, J., and Leadbeater, B. S. C. Eds. *The haptophyte algae*. Clarendon Press, Oxford, pp. 209-228.
- Marin, B., Klinberg, M. & Melkonian, M. 1998. Phylogenetic relationships among the Cryptophyta: analyses of nuclear-encoded SSU rRNA sequences support the monophyly of extant plastid-containing lineages. *Protist* 149:265-276.
- Marin, B., Nowack, E. & Melkonian, M. 2005. A plastid in the making: evidence for a second primary endosymbiosis. *Protist* 156:425-432.
- Martin, W. & Herrmann, R. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol* 118:9-17.
- Martin, W., Stoebe, B., Goremykin, V., Hansmann, S., Hasegawa, M. & Kowallik, K. V. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393:162-165.
- Martin, W., Rujan, T., Richly, E. et al. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial and chloroplast genomes reveals plastid phylogeny and



- thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. USA* 99:12246-12251.
- Martin, W., Deusch, O., Stawski, N., Grunheit, N. & Goremykin, V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. *Trends Plant Sci.* 10:1360-1385.
- Matsuzaki, M., Misumi, O., Shin-i, T. et al. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* 428:653-657.
- McFadden, G. I., Gilson, P. & Douglas, S. 1994a. The photosynthetic endosymbiont in cryptomonad cells produces both chloroplast and cytoplasmic-type ribosomes. *Journal of Cell Science* 107:649-657.
- McFadden, G. I., Gilson, P. & Hill, D. 1994b. *Goniomonas*: rRNA sequences indicate that this phagotrophic flagellate is a close relative of the host component of cryptomonads. *Eur. J. Phycol.* 29:29-32.
- McInerney, J. O. 1998. GCUA: General Codon Usage Analysis. *Bioinformatics* 14:372-373.
- Medlin, L. K., Barker, G. L. A., Campbell, L., Green, J. C., Hayes, P. K., Marie, D., Wrieden, S. & Vulot, D. 1996. Genetic characterization of *Emiliania huxleyi* (Haptophyta). *J. Marine Syst.* 9:13-31.
- Medlin, L. K., Kooistra, W. H. C. F., Potter, D., Saunders, G. W. & Andersen, R. A. 1997. Phylogenetic relationships of the 'golden algae' (haptophytes, heterokonts, chromophytes) and their plastids. *Plant Syst. Evol.* 11:S187-S219.

- Merchant, S. 2006. Trace metal utilization in chloroplasts. *In* Wise, R., and Hooper, J. Eds. *The structure and function of plastids*. Springer, The Netherlands, pp. 199-218.
- Moreira, D., Le Guyader, H. & Philippe, H. 2000. The origin of red algae and the evolution of chloroplasts. *Nature* 405:69-72.
- Morse, D., Salois, P., Markovic, P. & Hastings, J. 1995. A nuclear-encoded form II Rubisco in dinoflagellates. *Science* 268:1622-1624.
- Murray, S., Jorgensen, M., Ho, S., Patterson, D. & Jermiin, L. 2005. Improving the analysis of dinoflagellate phylogeny based on rDNA. *Protist* 156:269-286.
- Nisbet, R., Koumandou, V., Barbrook, A. & Howe, C. 2004. Novel plastid gene minicircles in the dinoflagellate *Amphidinium operculatum*. *Gene* 331:141-147.
- Nozaki, H., Matsuzaki, M., Misumi, O. & Kuroiwa, H. 2003a. Phylogeny of plastids based on cladistic analysis of gene loss inferred from complete plastid genome sequences. *J. Mol. Biol.* 57:377-382.
- Nozaki, H., Matsuzaki, M., Takahara, M. et al. 2003b. The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids. *J. Mol. Evol.* 56:485-497.
- Nozaki, H., Matsuzaki, M., Misumi, O. et al. 2004. Cyanobacterial genes transmitted to the nucleus before divergence of red algae in the chromista. *J. Mol. Evol.* 59:103-113.

- Nugent, J. & Palmer, J. D. 1991. RNA-mediated transfer of the gene *coxII* from the mitochondrion to the nucleus during flowering plant evolution. *Cell* 66:473-481.
- Obornik, M. & Green, B. R. 2005. Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. *Mol. Biol. Evol.* 22:2343-2353.
- Oda, K., Yamato, K., Ohta, E. et al. 1992. Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA: a primitive form of plant mitochondrial genome. *J. Mol. Biol.* 223:1-7.
- Ohta, N., Sato, N., Nozaki, H. & Kuroiwa, T. 1997. Analysis of the cluster of ribosomal protein genes in the plastid genome of a unicellular red alga *Cyanidioschyzon merolae*: translocation of the *str* cluster as an early event in the Rhodophyte-Chromophyte lineage of plastid evolution. *J. Mol. Evol.* 45:688-695.
- Ohta, N., Sato, N. & Kuroiwa, T. 1998. Structure and organization of the mitochondrial genome of the unicellular red alga *Cyanidioschyzon merolae* deduced from the complete nucleotide sequence. *Nucleic Acids Res.* 26:5190-5198.
- Ohta, N., Matsuzaki, M., Misumi, O., Miyagishima, S.-y., Nozaki, H., Tanaka, K., Shin-i, T., Kohara, Y. & Kuroiwa, T. 2003. Complete sequence and analysis of the plastid genome of the unicellular red alga *Cyanidioschyzon merolae*. *DNA Res.* 10:67-77.
- Okamoto, O. & Hastings, J. 2003. Genome-wide analysis of redox-regulated genes in a dinoflagellate. *Gene* 321:73-81.

- Paasche, E. 2002. A review of the coccolithophorid *Emiliana huxleyi* (Prymnesiophyceae), with particular reference to growth, coccolith formation, and calcification-photosynthesis interactions. *Phycologia* 40:503-529.
- Palmer, J. D. & Delwiche, C. F. 1996. Second-hand chloroplasts and the case of the disappearing nucleus. *Proc. Natl. Acad. Sci. USA* 93:7432-7435.
- Palmer, J. D. & Delwiche, C. F. 1998. The origin and evolution of plastids and their genomes. In Soltis, D., Soltis, P., and Doyle, J. J. Eds. *Molecular Systematics of Plants*. Kluger Academic Publishers, Norwell, Massachusetts, pp. 374-409.
- Palmer, J. D. 2003. The symbiotic birth and spread of plastids; how many times and whodunit? *J. Phycol.* 39:4-11.
- Pascher, A. 1910. Chrysomonaden aus dem Hirschberger Grossteiche. Untersuchungen über die Flora des Hirschberger Grossteiches. I. Teil. *Monogr. Abh. Int. Rev. Gesamten Hydrobiol. Hydrogr.* 1:1-66.
- Patron, N., Rogers, M. & Keeling, P. J. 2004. Gene replacement of fructose 1,6 biphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates. *Eukaryot. Cell* 3:1169-1175.
- Patron, N., Waller, R., Archibald, J. M. & Keeling, P. J. 2005. Complex protein targeting to dinoflagellate plastids. *J. Mol. Biol.* 348:1015-1024.
- Patron, N., Waller, R. & Keeling, P. J. 2006. A tertiary plastid uses genes from two endosymbionts. *J. Mol. Biol.* (in press).
- Patterson, D. 1999. The diversity of eukaryotes. *Am. Nat.* 65:S96-S124.
- Pfannschmidt, T., Nilsson, A. & Allen, J. 1999. Photosynthetic control of chloroplast gene expression. *Nature* 397:625-628.

- Philippe, H., Snell, E., Baptiste, E., Lopez, P., Holland, P. & Casane, D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol. Biol. Evol.* 21:1740-1752.
- Phillips, M., Delsuc, F. & Penny, D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* 21:1455-1458.
- Posada, D. & Crandall, K. 1998. ModelTest: testing the model of DNA substitution. *Bioinformatics* 14:817-818.
- Pritchard, A. E., Seilhamer, J. J., Mahalingam, R., Sable, C. L., Venuti, S. E. & Cummings, D. J. 1990. Nucleotide sequence of the mitochondrial genome of *Paramecium*. *Nucleic Acids Res.* 18:173-180.
- Race, H., Herrmann, R. & Martin, W. 1999. Why have organelles retained genomes? *Trends Genet.* 15:364-370.
- Rae, P. & Spear, B. 1978. Macronuclear DNA of the hypotrichous ciliate *Oxytricha fallax*. *Proc. Natl Acad. Sci. USA* 75:4992-4996.
- Ralph, S. A., van Dooren, G., Waller, R., Crawford, M., Fraunholz, M. J., Foth, B. J., Tonkin, C. J., Roos, D. S. & McFadden, G. I. 2004. Metabolic maps and functions of the *Plasmodium falciparum* apicoplast. *Nature Reviews Microbiology* 2:203-216.
- Reith, M. & Munholland, J. 1995. Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. *Plant Mol. Biol. Rep.* 13:333-335.
- Richard, O., Bonnard, G., Grienemberger, J. M., Kloareg, B. & Boyen, C. 1998. Transcription initiation and RNA processing in the mitochondria of the red

- alga *Chondrus crispus*: Convergence in the evolution of transcription mechanisms in mitochondria. *J. Mol. Biol.* 283:549-557.
- Riebesell, U., Zondervan, I., Rost, B., Tortell, P. D., Zeebe, R. E. & Morel, F. M. M. 2000. Reduced calcification of marine plankton in response to increased atmospheric CO<sub>2</sub>. *Nature* 407:364-367.
- Rivera, M. & Lake, J. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431:152-155.
- Rizzo, P. 1987. Biochemistry of the dinoflagellate nucleus. In Taylor, F. Ed *The biology of dinoflagellates*. Blackwell Scientific Publishing, Oxford, pp. 143-173.
- Rodriguez-Ezpeleta, N., Brinkmann, H., Burey, S., Roure, B., Burger, G., Löffelhardt, W., Bohnert, H., Phillipe, H. & Lang, B. F. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr. Biol.* 15:1325-1330.
- Roger, A. J. 1999. Reconstructing early events in eukaryotic evolution. *Am. Nat.* 154:S146-S163.
- Rogers, S., Rogers, M., Saunders, G. W. & Holt, G. 1986. Isolation of mutants sensitive to 2-aminopurine and alkylating agents and evidence for the role of DNA methylation in *Penicillium chrysogenum*. *Curr. Genet.* 10:557-560.
- Rohmer, M., Knani, M., Simonin, P., Sutter, B. & Sahm, H. 1993. Isoprenoid biosynthesis in bacteria: a novel pathway for the early steps leading to isopentenyl diphosphate. *Biochem. J.* 295:517-524.

- Saez, A., Probert, I., Young, J., Edvardsen, B., Eikrem, W. & Medlin, L. K. 2004. A review of the phylogeny of the Haptophyta. *In* Thierstein, H., and Young, J. Eds. *Coccolithophores. From molecular processes to global impact*. Springer-Verlag, Berlin, pp. 251-269.
- Saldarriaga, J. F., Taylor, F., Keeling, P. J. & Cavalier-Smith, T. 2001. Dinoflagellate Nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements. *J. Mol. Evol.* 53:204-213.
- Saldarriaga, J. F., McEwan, M., Fast, N. M., Taylor, F. & Keeling, P. J. 2003. Multiple protein phylogenies show that *Oxyrrhis marina* and *Perkinsus marinus* are early branches of the dinoflagellate lineage. *Int. J. Syst. Evol. Microbiol.* 53:355-365.
- Sambrook, J. & Russell, D. W. 2001. Molecular cloning. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York., pp.
- Sanchez-Puerta, M. V., Bachvaroff, T. R. & Delwiche, C. F. 2004. The complete mitochondrial genome sequence of the haptophyte *Emiliania huxleyi* and its relation to heterokonts. *DNA Res.* 11:1-10.
- Sanchez-Puerta, M. V., Bachvaroff, T. R. & Delwiche, C. F. 2005. The complete plastid genome sequence of the haptophyte *Emiliania huxleyi*: a comparison to other plastid genomes. *DNA Res.* 12:151-156.
- Saunders, G. W., Hill, D., Sexton, J. P. & Andersen, R. A. 1997. Small subunit ribosomal RNA sequences from selected dinoflagellates: testing classical evolutionary hypotheses with molecular systematic methods. *Pl. Syst. Evol.* S11:237-259.

- Schimodaira, H. & Hasegawa, M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246-1247.
- Schimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51:492-508.
- Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. 2002. Tree-Puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502-504.
- Schnepf, E. & Elbrächter, M. 1992. Nutritional strategies in dinoflagellates. *Eur. J. Phycol.* 28:3-24.
- Schnepf, E. & Elbrächter, M. 1999. Dinophyte chloroplasts and phylogeny - a review. *Grana* 38:81-97.
- Schopf, J., Kudryavtsev, A., Agresti, D., Wdowiak, T. & Czaja, A. 2002. Laser-Raman imagery of Earth's earliest fossils. *Nature* 416:73-76.
- Sekiguchi, H., Moriya, M., Nakayama, T. & Inouye, I. 2002. Vestigial chloroplasts in heterotrophic stramenopiles *Pteridomonas danica* and *Ciliophrys infusionum* (Dictyochophyceae). *Protist* 153:157-167.
- Shimada, H. & Sugiura, C. 1991. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. *Nucleic Acids Res.* 19:983-995.
- Shivji, M. 1991. Organization of the chloroplast genome in the red alga *Porphyra yezoensis*. *Curr. Genet.* 19:49-54.



- Sijtsma, L. & Swaaf, M. 2004. Biotechnological production and applications of the omega-3 polyunsaturated fatty acid docosahexaenoic acid. *Appl. Microbiol. Biotechnol.* 64:146-153.
- Simpson, A. & Roger, A. J. 2004. The real kingdoms of eukaryotes. *Curr. Biol.* 14:693-696.
- Stechmann, A. & Cavalier-Smith, T. 2003. Phylogenetic analysis of eukaryotes using heat-shock protein *Hsp90*. *J Mol Evol* 57:408-419.
- Steiner, J. & Löffelhardt, W. 2002. Protein import into cyanelles. *Trends in Plant Science* 7:72-77.
- Stiller, J. W., Reel, D. & Johnson, J. 2003. A single origin of plastids revisited: convergent evolution in organellar genome content. *J. Phycol.* 39:95-105.
- Stoebe, B. & Kowallik, K. V. 1999. Gene-cluster analysis in chloroplast genomics. *Trends Genet.* 15:344-347.
- Sugiura, C., Kobayashi, Y., Aoki, S., Sugita, C. & Sugita, M. 2003. Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA* from the chloroplast to the nucleus. *Nucleic Acids Res.* 31:5324-5331.
- Sugiura, M. 1995. The chloroplast genome. *Essays Biochem.* 30:49-57.
- Swofford, D., Olsen, G., Waddell, P. & Hillis, D. 2002. PAUP\*. Phylogenetic Analysis Using Parsimony (\*and other methods). Version 4. Sinauer Associates, Sunderland.
- Takishita, K., Nakano, K. & Uchida, A. 2000. Origin of the plastid in the anomalously pigmented dinoflagellate *Gymnodinium mikimotoi*

- (Gymnodiniales, Dinophyta) as inferred from phylogenetic analysis based on the gene encoding the large subunit of form I-type RuBisCo. *Phycol Res* 48:85-89.
- Takishita, K., Ishikura, M., Koike, K. & Maruyama, T. 2003. Comparison of phylogenies based on nuclear-encoded SSU rDNA and plastid-encoded *psbA* in the symbiotic dinoflagellate genus *Symbiodinium*. *Phycologia* 42:285-291.
- Tanikawa, N., Akimoto, H., Ogoh, K., Chun, W. & Ohmiya, Y. 2004. Expressed sequence tag analysis of the dinoflagellate *Lingulodinium polyedrum* during dark phase. *Photochemistry and Photobiology* 80:31-35.
- Taylor, F. 1987. Dinoflagellate morphology. In Taylor, F. Ed *The biology of dinoflagellates*. Blackwell Scientific Publications, Oxford, pp. 24-91.
- Taylor, F. 1999. Ultrastructure as a control for protistan molecular phylogeny. *Am. Nat.* 154:S125-S136.
- Tengs, T., Dahlberg, O. J., Shalchian-Tabrizi, K., Klaveness, D., Rudi, K., Delwiche, C. F. & Jakobsen, K. S. 2000. Phylogenetic analyses indicate that the 19'hexanoyloxy-fucoxanthin-containing dinoflagellates have tertiary plastids of haptophyte origin. *Mol. Biol. Evol.* 17:718-729.
- Tengs, T., Bowers, H. A., Ziman, A. P., Stoecker, D. K. & Oldach, D. W. 2001. Genetic polymorphism in *Gymnodinium galatheanum* chloroplast DNA sequences and development of a molecular detection assay. *Molecular Ecology* 10:515-523.
- Thatcher, T. & Gorovsky, M. 1994. Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. *Nucleic Acids Res.* 22:174-179.

- Timmis, J. N., Ayliffe, M. A. & Martin, W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews* 5:123-129.
- Turmel, M., Otis, C. & Lemieux, C. 2002. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. *Proc. Natl. Acad. Sci. USA* 99:11275-11280.
- Tuttle, R. & Loeblich, A. I. 1975. An optimal growth medium for the dinoflagellate *Cryptothecodinium cohnii*. *Phycologia* 14:1-8.
- Unsold, M., Marienfeld, J. R., Brandt, P. & Brennicke, A. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nature Genet.* 15:57-61.
- Van de Peer, Y. & De Wachter, R. 1997. Evolutionary relationships among eukaryotic crown taxa taking into account site-to-site variation in 18S rRNA. *J. Mol. Evol.* 45:619-630.
- van den Hoek, C., Mann, D. & Jahns, H. 1995. Algae. An introduction to phycology. Cambridge University Press, Cambridge, 623 pp.
- Van der Auwera, G., Hofmann, C., De Rijk, P. & Wachter, R. D. 1998. The origin of red algae and cryptomonad nucleomorphs: a comparative phylogeny based on small and large subunit rRNA sequences of *Palmaria palmata*, *Gracilaria verrucosa*, and the *Guillardia theta* nucleomorph. *Mol. Phylogenet. Evol.* 10:333-342.

- van Dooren, G. G., Schwartzbach, S. D., Osafune, T. & McFadden, G. I. 2001.  
Translocation of proteins across the multiple membranes of complex plastids.  
*Biochimica et Biophysica Acta* 1541:34-53.
- Wahlund, T., Hadaegh, A., Clark, R., Nguyen, B., Fanelli, M. & Read, B. 2004.  
Analysis of expressed sequence tags from calcifying cells of marine  
coccolithophorid (*Emiliana huxleyi*). *Mar. Biotechnol.* 6:278-290.
- Wakasugi, T., Nagai, T., Kapoor, M. et al. 1997. Complete nucleotide sequence of the  
chloroplast genome from the green alga *Chlorella vulgaris*: the existence of  
genes possibly involved in chloroplast division. *Proc. Natl. Acad. Sci. USA*  
94:5967-5972.
- Walsh, D. & Mann, S. 1995. Fabrication of hollow porous shells of calcium carbonate  
from self-organizing media. *Nature* 377:320-323.
- Wang, S.-L., Liu, X.-Q. & Douglas, S. 1997. The large ribosomal protein gene cluster  
of a cryptomonad plastid: gene organization, sequence and evolutionary  
implications. *Biochem. Mol. Biol. Int.* 41:1035-1044.
- Wang, Y. & Morse, D. 2006. The plastid-encoded *psbA* gene in the dinoflagellate  
*Gonyaulax* is not encoded on a minicircle. *Gene* (in press).
- Watson, G. & Tabita, F. 1997. Microbial ribulose 1,5-bisphosphate  
carboxylase/oxygenase: a molecule for phylogenetic and enzymological  
investigation. *FEMS Microbiology Letters* 146:13-22.
- Weber, A., Linka, M. & Bhattacharya, D. 2006. Single, ancient origin of a plastid  
metabolite translocator family in Plantae from an endomembrane-derived  
ancestor. *Eukaryot. Cell* 5:609-612.

- Wilson, R. 2005. Parasite plastids: approaching the end game. *Biol. Rev.* 80:129-153.
- Wilson, R. J. M. I. 2002. Progress with parasite plastids. *J. Mol. Biol.* 319:257-274.
- Winter, A. & Siesser, W. G. 1994. Coccolithophores. Cambridge University Press, Cambridge., 242 pp.
- Wissinger, B., Schuster, W. & Brennicke, A. 1991. *Trans*-splicing in *Oenothera* mitochondria: *nad1* mRNAs are edited in exon and *Trans*-splicing group II intron sequences. *Cell* 65:473-482.
- Woese, C. & Fox, G. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* 74:5088-5090.
- Woese, C., Kandler, O. & Wheelis, M. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* 87:4576-4579.
- Wolf, P. G., Rowe, C. A., Sinclair, R. B. & Hasebe, M. 2003. Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. *DNA Res.* 10:59-63.
- Wolfe, A. & dePamphilis, C. W. 1998. The effect of relaxed functional constraints on the photosynthetic gene *rbcL* in photosynthetic and nonphotosynthetic parasitic plants. *Mol. Biol. Evol.* 15:1243-1258.
- Wolfe, K., Morden, C. W. & Palmer, J. D. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc. Natl. Acad. Sci. USA* 89:10648-10652.

- Wong, J., New, D., Wong, J. & Hung, V. 2003. Histone-like proteins of the dinoflagellate *Cryptothecodinium cohnii* have homologies to bacterial DNA-binding proteins. *Eukaryot. Cell* 2:646-650.
- Wyman, S., Jansen, R. & Boore, J. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20:3252-3255.
- Xiao, S., Zhang, Y. & Knoll, A. 1998. Three-dimensional preservation of algae and animal embryos in a Neoproterozoic phosphorite. *Nature* 391:553-558.
- Xu, X., Adams, S., Chua, N. & Moller, S. 2005. AtNAP1 represents an atypical *sufB* protein in *Arabidopsis* plastids. *J. Biol. Chem.* 280:6648-6654.
- Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13:555-556.
- Yates, T., Salazar-Bravo, J. & Dragoo, J. 2004. The importance of the Tree of Life to society. In Cracraft, J., and Donoghue, M. Eds. *Assembling the Tree of Life*. Oxford University Press, New York, pp. 7-17.
- Yoon, H. S., Hackett, J. D. & Bhattacharya, D. 2002a. A single origin of the peridinin- and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis. *Proc. Natl. Acad. Sci. USA* 99:11724-11729.
- Yoon, H. S., Hackett, J. D., Pinto, G. & Bhattacharya, D. 2002b. The single, ancient origin of chromist plastids. *Proc. Natl. Acad. Sci. USA* 99:15507-15512.
- Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. & Bhattacharya, D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.* 21:809-818.

- Yoon, H. S., Hackett, J. D., Van Dolah, F., Nosenko, T., Lidie, K. & Bhattacharya, D. 2005. Tertiary endosymbiosis driven genome evolution in dinoflagellate algae. *Mol. Biol. Evol.* 22:1299-1308.
- Young, J., Didymus, J. M., Bown, P. R., Prins, B. & Mann, S. 1992. Crystal assembly and phylogenetic evolution in heterococcoliths. *Nature* 356:516-518.
- Zhang, H., Bhattacharya, D. & Lin, S. 2005. Phylogeny of dinoflagellates based on mitochondrial cytochrome b and nuclear small subunit rDNA sequence comparisons. *J. Phycol.* 41:411-420.
- Zhang, Z., Green, B. R. & Cavalier-Smith, T. 1999. Single gene circles in dinoflagellate chloroplast genomes. *Nature* 400:155-159.
- Zhang, Z., Green, B. R. & Cavalier-Smith, T. 2000. Phylogeny of ultra-rapidly evolving dinoflagellate chloroplast genes: a possible common origin for sporozoan and dinoflagellate plastids. *J. Mol. Evol.* 51:26-40.
- Zhang, Z., Cavalier-Smith, T. & Green, B. R. 2002. Evolution of dinoflagellate unigenic minicircles and the partially concerted divergence of their putative replicon origins. *Mol. Biol. Evol.* 19:489-500.
- Ziveri, P., Baumann, K., Bockel, B., Bollman, J. & Young, J. 2004. Biogeography of selected Holocene coccoliths in the Atlantic ocean. In Thierstein, H., and Young, J. Eds. *Coccolithophores. From molecular processes to global impact*. Springer-Verlag, Berlin, pp. 403-428.